PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12N

A2

(11) International Publication Number: **WO 98/59034**

(43) International Publication Date: 30 December 1998 (30.12.98)

(21) International Application Number: PCT/US98/13041

(22) International Filing Date: 23 June 1998 (23.06.98)

(30) Priority Data:
60/050,667 24 June 1997 (24.06.97) US

(71) Applicant *(for all designated States except US)*: HUMAN GENOME SCIENCES, INC. [US/US]; 9410 Key West Avenue, Rockville, MD 20850 (US).

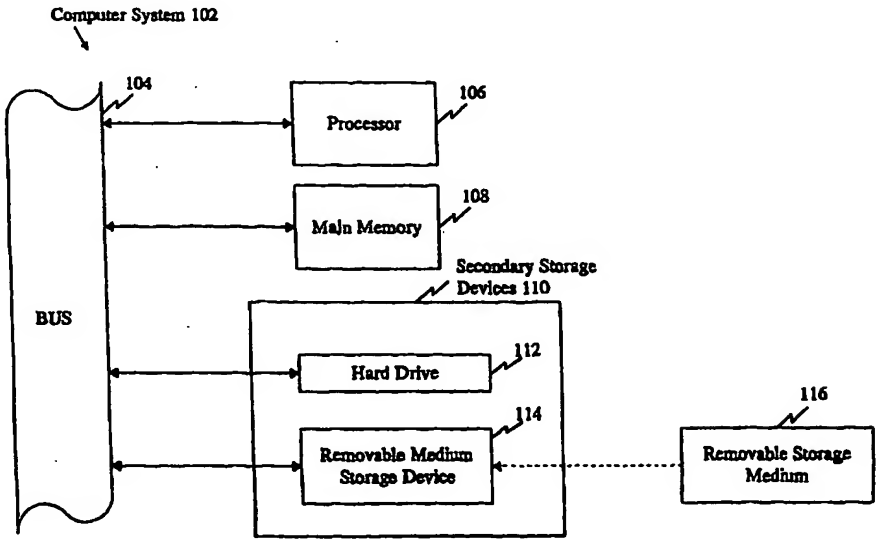
(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: FRASER, Claire, M. [US/US]; 11915 Glen Mill Road, Potomac, MD 20854 (US).

(74) Agents: BROOKES, A., Anders et al.; Human Genome Sciences, Inc., 9410 Key West Avenue, Rockville, MD 20850 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, GW, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: *TREPONEMA PALLIDUM* POLYNUCLEOTIDES AND SEQUENCES



(57) Abstract

The present invention provides polynucleotide sequences of the genome of *T. pallidum*, polypeptide sequences encoded by the polynucleotide sequences, corresponding polynucleotides and polypeptides, vectors and hosts comprising the polynucleotides, and assays and other uses thereof. The present invention further provides polynucleotide and polypeptide sequence information stored on computer readable media, and computer–based systems and methods which facilitate its use.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

Treponema pallidum Polynucleotides and Sequences

FIELD OF THE INVENTION

The present invention relates to the field of molecular biology. In particular, it relates to, among other things, nucleotide sequences of *Treponema pallidum*, contigs, ORFs, fragments, probes, primers and related polynucleotides thereof, peptides and polypeptides encoded by the sequences, and uses of the polynucleotides and sequences thereof, such as in fermentation, polypeptide production, assays and pharmaceutical development, among others.

BACKGROUND OF THE INVENTION

Spirochetes are a family of motile, unicellular, spiral-shaped bacteria which share a number of structural characteristics. Three genera of the spirochetes are pathogenic in humans: (a) *Treponema*, which includes the pathogens that cause syphilis (*T. pallidum*), yaws (*T. pertenue*), and pinta (*T. carateum*); (b) *Borrelia*, which includes the pathogens that cause epidemic and endemic relapsing fever and Lyme disease; and (c) *Leptospira*, which includes a wide variety of small spirochetes that cause mild to serious systemic human illness (Koff, A. B. and Rosen, T. *J. Am. Acad. Dermatol.* **29**:519-535 (1993)). In 1986, more than 27,000 cases of early infectious syphilis were diagnosed in the United States alone. Such statistics indicate that infection with *T. pallidum* is the largest source of human disease resulting from the spirochetes.

T. pallidum is morphologically indistinguishable from several other pathogenic spirochetes, but, in general, treponemes and other spirochetes, are easily identifiable when compared to other bacteria. A key morphological characteristic of *T. pallidum*, and other spirochetes, is the presence of a central protoplasmic cylinder composed primarily of peptidoglycan and one or more adjacent axial fibrils (also designated periplasmic flagella or endoflagella; Charon, N. W., *et al., Res. Microbiol.* **143**:597-603 (1992)). These structures provide a source of corkscrew-like motion to the treponemes. In aqueous media, treponemes move in an apparently random fashion and, unlike the majority of motile bacteria, continue to move in a more viscous medium. In tissues, treponemes are highly moldable to intercellular spaces; a characteristic which is thought to be mediated by the interactions of bacterial adhesins and cellular fibronectins.

Syphilis is the primary clinical manifestation of infection with *T. pallidum*. The clinical manifestations of syphilis can resemble many diseases. Syphilis is typically transmitted by sexual contact, but can also be transmitted transplacentally. The infecting organism multiplies at the site of infection within 10 to 60 days postinfection and results in a primary ulcer-like lesion termed a chancre. A small number of organisms move from the primary lesion to the regional lymph nodes and establish small infectious centers termed satellite buboes. Organisms from

these locations enter the blood stream and result in a systemic infection (Goens, J. L., *et al., Am. Fam. Physician* **50:**1013-1020 (1994)).

The secondary stage of syphilis manifests itself as a widespread skin rash and begins between two and twelve weeks following the primary infection. During this stage, the infected individual often experiences a low grade fever coupled with swollen lymph nodes. Also during this period, lesions of various degrees of severity may develop in a number of phyical locations including bone, liver, kidney, central nervous system (CNS), and other organs (Veeravahu, M. *Arch. Intern. Med.* **145:**132-134 (1985)). Such secondary infections are highly infectious, but will, in time, subside spontaneously.

A third stage of syphilis occurs in approximately 30% of infected, but not treated, individuals. The third stage occurs several years following the first and second stages. The lesions which characterize the third stage of infection are minor in terms of the number of organisms, but may be severe in terms of tissue damage. Such lesions may result in necrosis, scar formation, general paresis, damage to aortic valves, permanent blindness, and other extensive tissue damage, all probably related to a delayed type hypersensitivity reaction by the host to the *T. pallidum* organisms (Scheck, D. N. and Hook, E. W. 3rd *Infect. Dis. Clin. North Am.* **8:**769-795 (1994)).

A further, and increasingly common, complication of syphilis infection is coinfection with the human immunodeficiency virus (HIV). In fact, a recent study indicates that ulcerous genital diseases such as those exhibited during the primary stages of infection with syphilis may facilitate the transmission of HIV (Rufli, T. *Dermatologica* **179:**113-117 (1989)). In addition, it is clear that the CNS is regularly involved in the early stages of syphilis. In the timespan between the introduction of penecillin and other antibotics and the spread of HIV, early neurosyphilis was an exceptionally uncommon development. However, since the standard antibiotic dosage used to treat syphilis is not exceptionally high and since a successful treatment requires an adequate host immune response, individuals infected with HIV often exhibit a highly increased occurance of many neurosyphilis-related sequalae including asymptomatic neurosyphilis, syphilitic meinigitis, cranial nerve abnormalities, or cerebrovascular problems (Musher, D. M., *et al., Ann. Intern. Med.* **113:**872-881 (1990)).

T. pallidum has a remarkable ability to evade both the humoral and cellular components of the immune system. It was originally thought that the ability of *T. pallidum* to evade the immune system of the host organism was due to the presence of an outer coat of mucopolysaccharides. However, recent evidence suggests it is more likely that *T. pallidum* make use of the organization of the relative immunogenicity of its complement of outer membrane proteins to evade the immune system (Radolf, J. D. *Mol. Microbiol.* **16:**1067-1073 (1995)). Unlike most other bacterial outer membranes characterized thus far, the *T. pallidum* outer membrane contains a scarcity of immunogenic transmembrane proteins (with regard to *T. pallidum*, these are termed "rare outer membrane proteins"). Among the highly immunogenic proteins of treponemes are a number of lipoproteins anchored to the periplasmic leaflet of the cytoplasmic membrane. As a

result of their physical location, the lipoproteins may be less susceptible to typical immunologic surveillance (Norris, J. *Microbiol. Rev.* 57:750-779 (1993)). In addition to the periplasmic lipoproteins, *T. pallidum* also secretes a number of small, but immunogenic proteins which may induce an immune response (Hindersson, P. *et al., Res. Microbiol.* 143:629-639 (1992)).

It is clear that the etiology of diseases mediated or exacerbated by *T. pallidum* genes, and that characterizing the genes and their patterns of expression would add dramatically to our understanding of the organism and its host interactions. Knowledge of *T. pallidum* genes and genomic organization would dramatically improve understanding of disease etiology and lead to improved and new ways of preventing, ameliorating, arresting and reversing diseases. Moreover, characterized genes and genomic fragments of *T. pallidum* would provide reagents for, among other things, detecting, characterizing and controlling *T. pallidum* infections. There is a need therefore to characterize the genome of *T. pallidum* and for polynucleotides and sequences of this organism.

SUMMARY OF THE INVENTION

The present invention is based on the sequencing of fragments of the *T. pallidum* genome. The primary nucleotide sequences which were generated are provided in SEQ ID NOS:1-744.

The present invention provides the nucleotide sequence of several thousand contigs of the *T. pallidum* genome, which are listed in tables below and set out in the Sequence Listing submitted herewith, and representative fragments thereof, in a form which can be readily used, analyzed, and interpreted by a skilled artisan. In one embodiment, the present invention is provided as contiguous strings of primary sequence information corresponding to the nucleotide sequences depicted in SEQ ID NOS: 1-744.

The present invention further provides nucleotide sequences which are at least 95% identical to the nucleotide sequences of SEQ ID NOS: 1-744.

The nucleotide sequence of SEQ ID NOS: 1-744 , a representative fragment thereof, or a nucleotide sequence which is at least 95% identical to the nucleotide sequence of SEQ ID NOS: 1-744 may be provided in a variety of mediums to facilitate its use. In one application of this embodiment, the sequences of the present invention are recorded on computer readable media. Such media includes, but is not limited to: magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media.

The present invention further provides systems, particularly computer-based systems which contain the sequence information herein described stored in a data storage means. Such systems are designed to identify commercially important fragments of the *T. pallidum* genome.

Another embodiment of the present invention is directed to fragments of the *T. pallidum* genome having particular structural or functional attributes. Such fragments of the *T. pallidum*

genome of the present invention include, but are not limited to, fragments which encode peptides, hereinafter referred to as open reading frames or ORFs, fragments which modulate the expression of an operably linked ORF, hereinafter referred to as expression modulating fragments or EMFs, and fragments which can be used to diagnose the presence of *T. pallidum* in a sample, hereinafter referred to as diagnostic fragments or DFs.

Each of the ORFs in fragments of the *T. pallidum* genome disclosed in Tables 1, 2 and 3, and the EMFs found 5' to the ORFs, can be used in numerous ways as polynucleotide reagents. For instance, the sequences can be used as diagnostic probes or amplification primers for detecting or determining the presence of a specific microbe in a sample, to selectively control gene expression in a host and in the production of polypeptides, such as polypeptides encoded by ORFs of the present invention, particular those polypeptides that have a pharmacological activity.

The present invention further includes recombinant constructs comprising one or more fragments of the *T. pallidum* genome of the present invention. The recombinant constructs of the present invention comprise vectors, such as a plasmid or viral vector, into which a fragment of the *T. pallidum* has been inserted.

The present invention further provides host cells containing any of the isolated fragments of the *T. pallidum* genome of the present invention. The host cells can be a higher eukaryotic host cell, such as a mammalian cell, a lower eukaryotic cell, such as a yeast cell, or a procaryotic cell such as a bacterial cell.

The present invention is further directed to isolated polypeptides and proteins encoded by ORFs of the present invention. A variety of methods, well known to those of skill in the art, routinely may be utilized to obtain any of the polypeptides and proteins of the present invention. For instance, polypeptides and proteins of the present invention having relatively short, simple amino acid sequences readily can be synthesized using commercially available automated peptide synthesizers. Polypeptides and proteins of the present invention also may be purified from bacterial cells which naturally produce the protein. Yet another alternative is to purify polypeptide and proteins of the present invention from cells which have been altered to express them.

The invention further provides methods of obtaining homologs of the fragments of the *T. pallidum* genome of the present invention and homologs of the proteins encoded by the ORFs of the present invention. Specifically, by using the nucleotide and amino acid sequences disclosed herein as a probe or as primers, and techniques such as PCR cloning and colony/plaque hybridization, one skilled in the art can obtain homologs.

The invention further provides antibodies which selectively bind polypeptides and proteins of the present invention. Such antibodies include both monoclonal and polyclonal antibodies.

The invention further provides hybridomas which produce the above-described antibodies. A hybridoma is an immortalized cell line which is capable of secreting a specific monoclonal antibody.

The present invention further provides methods of identifying test samples derived from cells which express one of the ORFs of the present invention, or a homolog thereof. Such methods comprise incubating a test sample with one or more of the antibodies of the present invention, or one or more of the DFs of the present invention, under conditions which allow a skilled artisan to determine if the sample contains the ORF or product produced therefrom.

In another embodiment of the present invention, kits are provided which contain the necessary reagents to carry out the above-described assays.

Specifically, the invention provides a compartmentalized kit to receive, in close confinement, one or more containers which comprises: (a) a first container comprising one of the antibodies, or one of the DFs of the present invention; and (b) one or more other containers comprising one or more of the following: wash reagents, reagents capable of detecting presence of bound antibodies or hybridized DFs.

Using the isolated proteins of the present invention, the present invention further provides methods of obtaining and identifying agents capable of binding to a polypeptide or protein encoded by one of the ORFs of the present invention. Specifically, such agents include, as further described below, antibodies, peptides, carbohydrates, pharmaceutical agents and the like. Such methods comprise steps of: (a)contacting an agent with an isolated protein encoded by one of the ORFs of the present invention; and (b)determining whether the agent binds to said protein.

The present genomic sequences of *T. pallidum* will be of great value to all laboratories working with this organism and for a variety of commercial purposes. Many fragments of the *T. pallidum* genome will be immediately identified by similarity searches against GenBank or protein databases and will be of immediate value to *T. pallidum* researchers and for immediate commercial value for the production of proteins or to control gene expression.

The methodology and technology for elucidating extensive genomic sequences of bacterial and other genomes has and will greatly enhance the ability to analyze and understand chromosomal organization. In particular, sequenced contigs and genomes will provide the models for developing tools for the analysis of chromosome structure and function, including the ability to identify genes within large segments of genomic DNA, the structure, position, and spacing of regulatory elements, the identification of genes with potential industrial applications, and the ability to do comparative genomic and molecular phylogeny.

DESCRIPTION OF THE FIGURES

FIGURE 1 is a block diagram of a computer system (102) that can be used to implement computer-based systems of present invention.

FIGURE 2 is a schematic diagram depicting the data flow and computer programs used to collect, assemble, edit and annotate the contigs of the *T. pallidum* genome of the present invention. Both Macintosh and Unix platforms are used to handle the AB 373 and 377 sequence

data files, largely as described in Kerlavage *et al.*, *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences*, 585, IEEE Computer Society Press, Washington D.C. (1993). Factura (AB) is a Macintosh program designed for automatic vector sequence removal and end-trimming of sequence files. The program Loadis runs on a Macintosh platform and parses the feature data extracted from the sequence files by Factura to the Unix based *T. pallidum* relational database. Assembly of contigs (and whole genome sequences) is accomplished by retrieving a specific set of sequence files and their associated features using Extrseq, a Unix utility for retrieving sequences from an SQL database. The resulting sequence file is processed to trim portions of the sequences with a high rate ambiguous nucleotides. The sequence files were assembled using TIGR Assembler, an assembly engine designed at The Institute for Genomic Research (TIGR) for rapid and accurate assembly of thousands of sequence fragments. The collection of contigs generated by the assembly step is loaded into the database with the lassie program. Identification of open reading frames (ORFs) is accomplished by processing contigs with zorf. The ORFs are searched against *T. pallidum* sequences from GenBank and against all protein sequences using the BLASTN and BLASTP programs (using default parameters), described in Altschul *et al.*, *J. Mol. Biol. 215:* 403-410 (1990). Results of the ORF determination and similarity searching steps were loaded into the database. As described below, some results of the determination and the searches are set out in Tables 1-3.

DETAILED DESCRIPTION OF ILLUSTRATIVE EMBODIMENTS

The present invention is based on the sequencing of fragments of the *T. pallidum* genome and analysis of the sequences. The primary nucleotide sequences generated by sequencing the fragments are provided in SEQ ID NOS: 1-744. As used herein, the "primary sequence" refers to the nucleotide sequence represented by the IUPAC nomenclature system.).

In addition to the aforementioned *T. pallidum* polynucleotide and polynucleotide sequences, the present invention provides the nucleotide sequences of SEQ ID NOS: 1-744, ORF IDs and ORFs within, or representative fragments thereof, in a form which can be readily used, analyzed, and interpreted by a skilled artisan.

As used herein, a "representative fragment of the nucleotide sequence depicted in SEQ ID NOS:1-744" refers to any portion of the SEQ ID NOS: 1-744 which is not presently represented within a publicly available database. Preferred representative fragments of the present invention are *T. pallidum* open reading frames (ORFs), expression modulating fragment (EMFs) and fragments which can be used to diagnose the presence of *T. pallidum* in sample (DFs). A non-limiting identification of preferred representative fragments is provided in Tables 1-3. As discussed in detail below, the information provided in SEQ ID NOS:1-744 and in Tables 1-3 together with routine cloning, synthesis, sequencing and assay methods will enable those skilled in the art to clone and sequence all "representative fragments" of interest, including open reading frames encoding a large variety of *T. pallidum* proteins.

The present invention is further directed to nucleic acid molecules encoding portions or fragments of the nucleotide sequences described herein. Fragments include portions of the nucleotide sequences of SEQ ID NOS:1-744, at least 10 contiguous nucleotides in length selected from any two integers, one of which representing a 5' nucleotide position and a second of which representing a 3' nucleotide position, where the first nucleotide for each nucleotide sequence in SEQ ID NOS:1-744 is position 1. That is, every combination of a 5' and 3' nucleotide position that a fragment at least 10 contiguous nucleotides in length could occupy is included in the invention. At least means a fragment may be 10 contiguous nucleotide bases in length or any integer between 10 and the length of an entire nucleotide sequence of SEQ ID NOS:1-744 minus 1. Therefore, included in the invention are contiguous fragments specified by any 5' and 3' nucleotide base positions of a nucleotide sequences of SEQ ID NOS:1-744 wherein the contiguous fragment is any integer between 10 and the length of an entire nucleotide sequence minus 1.

Further, the invention includes polynucleotides comprising fragments specified by size, in nucleotides, rather than by nucleotide positions. The invention includes any fragment size, in contiguous nucleotides, selected from integers between 10 and the length of an entire ORF ID, ORF, or SEQ ID NO:, minus 1. Preferred sizes of contiguous nucleotide fragments include 20 nucleotides, 30 nucleotides, 40 nucleotides, 50 nucleotides. Other preferred sizes of contiguous nucleotide fragments, which may be useful as diagnostic probes and primers, include fragments 50-300 nucleotides in length which include, as discussed above, fragment sizes representing each integer between 50-300. Larger fragments are also useful according to the present invention corresponding to most, if not all, of the nucleotide sequences shown in Tables 1-3 (ORF IDs) and SEQ ID NOS:1-744. The preferred sizes are, of course, meant to exemplify not limit the present invention as all size fragments, representing any integer between 10 and the length of an entire nucleotide sequence minus 1, of each ORF ID, ORF, and SEQ ID NO:, are included in the invention.

The present invention also provides for the exclusion of any fragment, specified by 5' and 3' base positions or by size in nucleotide bases as described above for any ORF ID or SEQ ID NOS:1-744. Any number of fragments of nucleotide sequences in ORF IDs or SEQ ID NOS:1-744, specified by 5' and 3' base positions or by size in nucleotides, as described above, may be excluded from the present invention.

While the presently disclosed sequences of SEQ ID NOS: 1-744 are highly accurate, sequencing techniques are not perfect and, in relatively rare instances, further investigation of a fragment or sequence of the invention may reveal a nucleotide sequence error present in a nucleotide sequence disclosed in SEQ ID NOS: 1-744. However, once the present invention is made available (*i.e.*, once the information in SEQ ID NOS: 1-744 and Tables 1-3 has been made available), resolving a rare sequencing error in SEQ ID NOS: 1-744 will be well within the skill of the art. The present disclosure makes available sufficient sequence information to allow any of the described contigs or portions thereof to be obtained readily by straightforward application of

routine techniques. Further sequencing of such polynucleotide may proceed in like manner using manual and automated sequencing methods which are employed ubiquitous in the art. Nucleotide sequence editing software is publicly available. For example, Applied Biosystem's (AB) AutoAssembler can be used as an aid during visual inspection of nucleotide sequences. By employing such routine techniques potential errors readily may be identified and the correct sequence then may be ascertained by targeting further sequencing effort, also of a routine nature, to the region containing the potential error.

Even if all of the very rare sequencing errors in SEQ ID NOS: 1-744 were corrected, the resulting nucleotide sequences would still be at least 95% identical, nearly all would be at least 99% identical, and the great majority would be at least 99.9% identical to the nucleotide sequences of SEQ ID NOS: 1-7441-744.

As discussed elsewhere herein, polynucleotides of the present invention readily may be obtained by routine application of well known and standard procedures for cloning and sequencing DNA. Detailed methods for obtaining libraries and for sequencing are provided below, for instance. A wide variety of *T. pallidum* strains can be used to prepare *T. pallidum* genomic DNA for cloning and for obtaining polynucleotides of the present invention which are known in th art.

The nucleotide sequences of the genomes from different strains of *T. pallidum* differ somewhat. However, the nucleotide sequences of the genomes of all *T. pallidum* strains will be at least 95% identical, in corresponding part, to the nucleotide sequences provided in SEQ ID NOS: 1-744 and the ORF IDs and ORFs within. Nearly all will be at least 99% identical and the great majority will be 99.9% identical.

The present application is further directed to nucleic acid molecules at least 90%, 95%, 96%, 97%, 98% or 99% identical to a nucleic acid sequence shown in SEQ ID NOS: 1-744, the ORF IDs and ORFs within. The above nucleic acid sequences are included irrespective of whether they encode a polypeptide having *T. pallidum* activity. This is because even where a particular nucleic acid molecule does not encode a polypeptide having *T. pallidum* activity, one of skill in the art would still know how to use the nucleic acid molecule, for instance, as a hybridization probe. Uses of the nucleic acid molecules of the present invention that do not encode a polypeptide having *T. pallidum* activity include, *inter alia*, isolating an *T. pallidum* gene or allelic variants thereof from a DNA library, and detecting *T. pallidum* mRNA expression samples, environmental samples, suspected of containing *T. pallidum* by Northern Blot, PCR, or similar analysis.

Preferred, are nucleic acid molecules having sequences at least 90%, 95%, 96%, 97%, 98% or 99% identical to the nucleic acid sequence shown in SEQ ID NOS: 1-744, the ORF IDs, and the ORF within each ORF ID, which do, in fact, encode a polypeptide having *T. pallidum* protein activity By "a polypeptide having *T. pallidum* activity" is intended polypeptides exhibiting activity similar, but not necessarily identical, to an activity of the *T. pallidum* protein of the invention, as measured in a particular biological assay suitable for measuring activity of the

specified protein.

Due to the degeneracy of the genetic code, one of ordinary skill in the art will immediately recognize that a large number of the nucleic acid molecules having a sequence at least 90%, 95%, 96%, 97%, 98%, or 99% identical to the nucleic acid sequences shown in SEQ ID NOS: 1-744, the ORF IDs, and the ORF within each ORF ID, will encode a polypeptide having *T. pallidum* protein activity. In fact, since degenerate variants of these nucleotide sequences all encode the same polypeptide, this will be clear to the skilled artisan even without performing the above described comparison assay. It will be further recognized in the art that, for such nucleic acid molecules that are not degenerate variants, a reasonable number will also encode a polypeptide having *T. pallidum* protein activity. This is because the skilled artisan is fully aware of amino acid substitutions that are either less likely or not likely to significantly effect protein function (e.g., replacing one aliphatic amino acid with a second aliphatic amino acid), as further described below.

The biological activity or function of the polypeptides of the present invention are expected to be similar or identical to polypeptides from other bacteria that share a high degree of structural identity/similarity. Table 1-3 lists accession numbers and descriptions for the closest matching sequences of polypeptides available through Genbank. It is therefore expected that the biological activity or function of the polypeptides of the present invention will be similar or identical to those polypeptides from other bacterial genuses, species, or strains listed in Table 1-3.

By a polynucleotide having a nucleotide sequence at least, for example, 95% "identical" to a reference nucleotide sequence of the present invention, it is intended that the nucleotide sequence of the polynucleotide is identical to the reference sequence except that the polynucleotide sequence may include up to five point mutations per each 100 nucleotides of the reference nucleotide sequence encoding the *T. pallidum* polypeptide. In other words, to obtain a polynucleotide having a nucleotide sequence at least 95% identical to a reference nucleotide sequence, up to 5% of the nucleotides in the reference sequence may be deleted, inserted, or substituted with another nucleotide. The query sequence may be an entire sequence shown in SEQ ID NOS: 1-744, the ORF IDs, or the ORF within each ORF ID, or any fragment specified as described herein.

As a practical matter, whether any particular nucleic acid molecule or polypeptide is at least 90%, 95%, 96%, 97%, 98% or 99% identical to a nucleotide sequence of the presence invention can be determined conventionally using known computer programs. A preferred method for determining the best overall match between a query sequence (a sequence of the present invention) and a subject sequence, also referred to as a global sequence alignment, can be determined using the FASTDB computer program based on the algorithm of Brutlag et al. *See* Brutlag et al. (1990) Comp. App. Biosci. 6:237-245. In a sequence alignment the query and subject sequences are both DNA sequences. An RNA sequence can be compared by first converting U's to T's. The result of said global sequence alignment is in percent identity.

Preferred parameters used in a FASTDB alignment of DNA sequences to calculate percent identity are: Matrix=Unitary, k-tuple=4, Mismatch Penalty=1, Joining Penalty=30, Randomization Group Length=0, Cutoff Score=1, Gap Penalty=5, Gap Size Penalty 0.05, Window Size=500 or the lenght of the subject nucleotide sequence, whichever is shorter.

If the subject sequence is shorter than the query sequence because of 5' or 3' deletions, not because of internal deletions, a manual correction must be made to the results. This is because the FASTDB program does not account for 5' and 3' truncations of the subject sequence when calculating percent identity. For subject sequences truncated at the 5' or 3' ends, relative to the query sequence, the percent identity is corrected by calculating the number of bases of the query sequence that are 5' and 3' of the subject sequence, which are not matched/aligned, as a percent of the total bases of the query sequence. Whether a nucleotide is matched/aligned is determined by results of the FASTDB sequence alignment. This percentage is then subtracted from the percent identity, calculated by the above FASTDB program using the specified parameters, to arrive at a final percent identity score. This corrected score is what is used for the purposes of the present invention. Only nucleotides outside the 5' and 3' nucleotides of the subject sequence, as displayed by the FASTDB alignment, which are not matched/aligned with the query sequence, are calculated for the purposes of manually adjusting the percent identity score.

For example, a 90 nucleotide subject sequence is aligned to a 100 nucleotide query sequence to determine percent identity. The deletions occur at the 5' end of the subject sequence and therefore, the FASTDB alignment does not show a matched/alignment of the first 10 nucleotides at 5' end. The 10 unpaired nucleotides represent 10% of the sequence (number of nucleotides at the 5' and 3' ends not matched/total number of nucleotides in the query sequence) so 10% is subtracted from the percent identity score calculated by the FASTDB program. If the remaining 90 nucleotides were perfectly matched the final percent identity would be 90%. In another example, a 90 nucleotide subject sequence is compared with a 100 nucleotide query sequence. This time the deletions are internal deletions so that there are no nucleotides on the 5' or 3' of the subject sequence which are not matched/aligned with the query. In this case the percent identity calculated by FASTDB is not manually corrected. Once again, only nucleotides 5' and 3' of the subject sequence which are not matched/aligned with the query sequence are manually corrected for. No other manual corrections are to made for the purposes of the present invention.

COMPUTER RELATED EMBODIMENTS

The nucleotide sequences provided in SEQ ID NOS: 1-744, including ORF IDs and corresponding ORFs, a representative fragment thereof, or a nucleotide sequence at least 95%, preferably at least 99% and most preferably at least 99.9% identical to said polynucleotide sequences may be "provided" in a variety of mediums to facilitate use thereof. As used herein, "provided" refers to a manufacture, other than an isolated nucleic acid molecule, which contains a nucleotide sequence of the present invention. Such a manufacture provides a large portion of the

T. pallidum genome and parts thereof (*e.g.*, a *T. pallidum* open reading frame (ORF)) in a form which allows a skilled artisan to examine the manufacture using means not directly applicable to examining the *T. pallidum* genome or a subset thereof as it exists in nature or in purified form.

In one application of this embodiment, a nucleotide sequence of the present invention can be recorded on computer readable media. As used herein, "computer readable media" refers to any medium which can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD- ROM; electrical storage media such as RAM and ROM; and hybrids of these categories, such as magnetic/optical storage media. A skilled artisan can readily appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising computer readable medium having recorded thereon a nucleotide sequence of the present invention. Likewise, it will be clear to those of skill how additional computer readable media that may be developed also can be used to create analogous manufactures having recorded thereon a nucleotide sequence of the present invention.

As used herein, "recorded" refers to a process for storing information on computer readable medium. A skilled artisan can readily adopt any of the presently know methods for recording information on computer readable medium to generate manufactures comprising the nucleotide sequence information of the present invention.

A variety of data storage structures are available to a skilled artisan for creating a computer readable medium having recorded thereon a nucleotide sequence of the present invention. The choice of the data storage structure will generally be based on the means chosen to access the stored information. In addition, a variety of data processor programs and formats can be used to store the nucleotide sequence information of the present invention on computer readable medium. The sequence information can be represented in a word processing text file, formatted in commercially- available software such as WordPerfect and MicroSoft Word, or represented in the form of an ASCII file, stored in a database application, such as DB2, Sybase, Oracle, or the like. A skilled artisan can readily adapt any number of data-processor structuring formats (*e.g.*, text file or database) in order to obtain computer readable medium having recorded thereon the nucleotide sequence information of the present invention.

Computer software is publicly available which allows a skilled artisan to access sequence information provided in a computer readable medium. Thus, by providing in computer readable form the nucleotide sequences of SEQ ID NOS: 1-744, including ORF IDs and corresponding ORFs, a representative fragment thereof, or a nucleotide sequence at least 95%, preferably at least 99% and most preferably at least 99.9% identical to said polynucleotide sequences, the present invention enables the skilled artisan routinely to access the provided sequence information for a wide variety of purposes.

The examples which follow demonstrate how software which implements the BLAST (Altschul *et al.*, *J. Mol. Biol. 215*:403-410 (1990)) and BLAZE (Brutlag *et al.*, *Comp. Chem. 17*:203-207 (1993)) search algorithms on a Sybase system was used to identify open reading

frames (ORFs) within the *T. pallidum* genome which contain homology to ORFs or proteins from both *T. pallidum* and from other organisms. Among the ORFs discussed herein are protein encoding fragments of the *T. pallidum* genome useful in producing commercially important proteins, such as enzymes used in fermentation reactions and in the production of commercially useful metabolites.

The present invention further provides systems, particularly computer-based systems, which contain the sequence information described herein. Such systems are designed to identify, among other things, commercially important fragments of the *T. pallidum* genome.

As used herein, "a computer-based system" refers to the hardware means, software means, and data storage means used to analyze the nucleotide sequence information of the present invention. The minimum hardware means of the computer-based systems of the present invention comprises a central processing unit (CPU), input means, output means, and data storage means. A skilled artisan can readily appreciate that any one of the currently available computer-based system are suitable for use in the present invention.

As stated above, the computer-based systems of the present invention comprise a data storage means having stored therein a nucleotide sequence of the present invention and the necessary hardware means and software means for supporting and implementing a search means.

As used herein, "data storage means" refers to memory which can store nucleotide sequence information of the present invention, or a memory access means which can access manufactures having recorded thereon the nucleotide sequence information of the present invention.

As used herein, "search means" refers to one or more programs which are implemented on the computer- based system to compare a target sequence or target structural motif with the sequence information stored within the data storage means. Search means are used to identify fragments or regions of the present genomic sequences which match a particular target sequence or target motif. A variety of known algorithms are disclosed publicly and a variety of commercially available software for conducting search means are and can be used in the computer-based systems of the present invention. Examples of such software includes, but is not limited to, MacPattern (EMBL), BLASTN and BLASTX (NCBIA). A skilled artisan can readily recognize that any one of the available algorithms or implementing software packages for conducting homology searches can be adapted for use in the present computer-based systems.

As used herein, a "target sequence" can be any DNA or amino acid sequence of six or more nucleotides or two or more amino acids. A skilled artisan can readily recognize that the longer a target sequence is, the less likely a target sequence will be present as a random occurrence in the database. The most preferred sequence length of a target sequence is from about 10 to 100 amino acids or from about 30 to 300 nucleotide residues. However, it is well recognized that searches for commercially important fragments, such as sequence fragments involved in gene expression and protein processing, may be of shorter length.

As used herein, "a target structural motif," or "target motif," refers to any rationally selected sequence or combination of sequences in which the sequence(s) are chosen based on a three-dimensional configuration which is formed upon the folding of the target motif. There are a variety of target motifs known in the art. Protein target motifs include, but are not limited to, enzymic active sites and signal sequences. Nucleic acid target motifs include, but are not limited to, promoter sequences, hairpin structures and inducible expression elements (protein binding sequences).

A variety of structural formats for the input and output means can be used to input and output the information in the computer-based systems of the present invention. A preferred format for an output means ranks fragments of the *T. pallidum* genomic sequences possessing varying degrees of homology to the target sequence or target motif. Such presentation provides a skilled artisan with a ranking of sequences which contain various amounts of the target sequence or target motif and identifies the degree of homology contained in the identified fragment.

A variety of comparing means can be used to compare a target sequence or target motif with the data storage means to identify sequence fragments of the *T. pallidum* genome. In the present examples, implementing software which implement the BLAST and BLAZE algorithms, described in Altschul *et al.*, *J. Mol. Biol. 215:* 403-410 (1990), is used to identify open reading frames within the *T. pallidum* genome. A skilled artisan can readily recognize that any one of the publicly available homology search programs can be used as the search means for the computer-based systems of the present invention. Of course, suitable proprietary systems that may be known to those of skill also may be employed in this regard.

Figure 1 provides a block diagram of a computer system illustrative of embodiments of this aspect of present invention. The computer system 102 includes a processor 106 connected to a bus 104. Also connected to the bus 104 are a main memory 108 (preferably implemented as random access memory, RAM) and a variety of secondary storage devices 110, such as a hard drive 112 and a removable medium storage device 114. The removable medium storage device 114 may represent, for example, a floppy disk drive, a CD-ROM drive, a magnetic tape drive, *etc.* A removable storage medium 116 (such as a floppy disk, a compact disk, a magnetic tape, *etc.*) containing control logic and/or data recorded therein may be inserted into the removable medium storage device 114. The computer system 102 includes appropriate software for reading the control logic and/or the data from the removable medium storage device 114, once it is inserted into the removable medium storage device 114.

A nucleotide sequence of the present invention may be stored in a well known manner in the main memory 108, any of the secondary storage devices 110, and/or a removable storage medium 116. During execution, software for accessing and processing the genomic sequence (such as search tools, comparing tools, *etc.*) reside in main memory 108, in accordance with the requirements and operating parameters of the operating system, the hardware system and the software program or programs.

BIOCHEMICAL EMBODIMENTS

Other embodiments of the present invention are directed to isolated fragments of the *T. pallidum* genome. The fragments of the *T. pallidum* genome of the present invention include, but are not limited to fragments which encode peptides, hereinafter open reading frames (ORFs), fragments which modulate the expression of an operably linked ORF, hereinafter expression modulating fragments (EMFs) and fragments which can be used to diagnose the presence of *T. pallidum* in a sample, hereinafter diagnostic fragments (DFs).

As used herein, an "isolated nucleic acid molecule" or an "isolated fragment of the *T. pallidum* genome" refers to a nucleic acid molecule possessing a specific nucleotide sequence which has been subjected to purification means to reduce, from the composition, the number of compounds which are normally associated with the composition. Particularly, the term refers to the nucleic acid molecules having the sequences set out in SEQ ID NOS: 1-744, to representative fragments thereof as described above including ORF IDs and ORFs, to polynucleotides at least 95%, preferably at least 96%, 97%, 98%, or 99% and especially preferably at least 99.9% identical in sequence thereto, also as set out above.

A variety of purification means can be used to generate the isolated fragments of the present invention. These include, but are not limited to methods which separate constituents of a solution based on charge, solubility, or size.

In one embodiment, *T. pallidum* DNA can be enzymatically sheared to produce fragments of 15-20 kb in length. These fragments can then be used to generate a *T. pallidum* library by inserting them into lambda clones as described in the Examples below. Primers flanking, for example, an ORF, such as those enumerated in the ORF IDs of Tables 1-3, can then be generated using nucleotide sequence information provided in SEQ ID NOS: 1-744. Well known and routine techniques of PCR cloning then can be used to isolate the ORF from the lambda DNA library or *T. pallidum* genomic DNA. Thus, given the availability of SEQ ID NOS:1-744, the information in Tables 1, 2 and 3, and the information that may be obtained readily by analysis of the sequences of SEQ ID NOS:1-744 using methods set out above, those of skill will be enabled by the present disclosure to isolate any ORF-containing or other nucleic acid fragment of the present invention.

The isolated nucleic acid molecules of the present invention include, but are not limited to single stranded and double stranded DNA, and single stranded RNA. For purposes of numbering and reference to polynucleotide and polypeptide sequences the entire sequence of each sequence of SEQ ID NOS:1-744 is included with the first nucleotide being position 1. Therefore, for reference purposes the numbering used in the present invention is that provided in the sequence listing for SEQ ID NOS:1-744.

As used herein, an open reading frame (ORF), means a series of nucleotide triplets coding for amino acid residues without any termination codons and is a sequence translatable into protein. Further, unless specified, the term "ORF" for each ORF ID is defined by the termination

codon at the 3' end and the 5' most methionine codon, at the 5' end, in frame with said 3' termination codon. Unless specified, the term "ORF" also refers to a particular polypeptide sequence defined by the ORF polynucleotide sequence, wherein the N-terminus is defined by the 5' most methionine codon in frame with the termination codon at the 3' end of the ORF ID and the C-terminus is defined by the last codon before the said 3' termination codon. As used herein, an ORF ID represents a sequence without any internal termination codons flanked by termination codons.

Tables 1, 2, and 3 list ORF IDs in the *T. pallidum* genomic contigs of the present invention that were identified as putative coding regions by the GeneMark software using organism-specific second-order Markov probability transition matrices. It will be appreciated that other criteria can be used, in accordance with well known analytical methods, such as those discussed herein, to generate more inclusive, more restrictive, or more selective lists.

Table 1 sets out ORF IDs in the *T. pallidum* contigs of the present invention that over a continuous region of at least 50 bases are 95% or more identical (by BLAST analysis) to a nucleotide sequence available through GenBank in June, 1997.

Table 2 sets out ORF IDs in the *T. pallidum* contigs of the present invention that are not in Table 1 and match, with a BLASTP probability score of 0.01 or less, a polypeptide sequence available through GenBank in July, 1996.

Table 3 sets out ORF IDs in the *T. pallidum* contigs of the present invention that do not match significantly, by BLASTP analysis, a polypeptide sequence available through GenBank in July, 1996.

In each table, the first and second columns identify the ORF ID by, respectively, contig number and ORF ID number within the contig; the third column indicates the first nucleotide of the ORF ID, counting from the 5' end of the contig strand; and the fourth column indicates the last nucleotide of the ORF ID, counting from the 5' end of the contig strand.

In Tables 1 and 2, column six, lists the Reference for the closest matching sequence available through GenBank. These reference numbers are the databases entry numbers commonly used by those of skill in the art, who will be familiar with their denominators. Descriptions of the nomenclature are available from the National Center for Biotechnology Information. Column seven in Tables 1 and 2 provides the gene name of the matching sequence; column eight provides the BLAST identity score from the comparison of the ORF and the homologous gene; and column nine indicates the length in nucleotides of the highest scoring segment pair identified by the BLAST identity analysis.

In Table 3, the last column, column six, indicates the length of each ORF ID in amino acid residues.

The concepts of percent identity and percent similarity of two polypeptide sequences is well understood in the art. For example, two polypeptides 10 amino acids in length which differ at three amino acid positions (*e.g.*, at positions 1, 3 and 5) are said to have a percent identity of 70%. However, the same two polypeptides would be deemed to have a percent similarity of

80% if, for example at position 5, the amino acids moieties, although not identical, were "similar" (*i.e.*, possessed similar biochemical characteristics). Many programs for analysis of nucleotide or amino acid sequence similarity, such as FASTA and BLAST specifically list percent identity of a matching region as an output parameter. Thus, for instance, Tables 1 and 2 herein enumerate the percent identity of the highest scoring segment pair in each ORF and its listed relative. Further details concerning the algorithms and criteria used for homology searches are provided below and are described in the pertinent literature highlighted by the citations provided below.

It will be appreciated that other criteria can be used to generate more inclusive and more exclusive listings of the types set out in the tables. As those of skill will appreciate, narrow and broad searches both are useful. Thus, a skilled artisan can readily identify ORFs in contigs of the *T. pallidum* genome other than those specified for Tables 1-3, such as ORFs which are overlapping or encoded by the opposite strand of an identified ORF in addition to those ascertainable using the computer-based systems of the present invention.

As used herein, an "expression modulating fragment," EMF, means a series of nucleotide molecules which modulates the expression of an operably linked ORF or EMF.

As used herein, a sequence is said to "modulate the expression of an operably linked sequence" when the expression of the sequence is altered by the presence of the EMF. EMFs include, but are not limited to, promoters, and promoter modulating sequences (inducible elements). One class of EMFs are fragments which induce the expression or an operably linked ORF in response to a specific regulatory factor or physiological event.

EMF sequences can be identified within the contigs of the *T. pallidum* genome by their proximity to the ORF IDs provided in Tables 1-3 and ORFs within each ORF ID. An intergenic segment, or a fragment of the intergenic segment, from about 10 to 200 nucleotides in length, taken from any one of the ORFs of Tables 1-3 will modulate the expression of an operably linked ORF in a fashion similar to that found with the naturally linked ORF sequence. As used herein, an "intergenic segment" refers to fragments of the *T. pallidum* genome which are between two ORF(s) herein described. EMFs also can be identified using known EMFs as a target sequence or target motif in the computer-based systems of the present invention. Further, the two methods can be combined and used together.

The presence and activity of an EMF can be confirmed using an EMF trap vector. An EMF trap vector contains a cloning site linked to a marker sequence. A marker sequence encodes an identifiable phenotype, such as antibiotic resistance or a complementing nutrition auxotrophic factor, which can be identified or assayed when the EMF trap vector is placed within an appropriate host under appropriate conditions. As described above, a EMF will modulate the expression of an operably linked marker sequence. A more detailed discussion of various marker sequences is provided below. A sequence which is suspected as being an EMF is cloned in all three reading frames in one or more restriction sites upstream from the marker sequence in the EMF trap vector. The vector is then transformed into an appropriate host using known

procedures and the phenotype of the transformed host in examined under appropriate conditions. As described above, an EMF will modulate the expression of an operably linked marker sequence.

As used herein, a "diagnostic fragment," DF, means a series of nucleotide molecules which selectively hybridize to *T. pallidum* sequences. DFs can be readily identified by identifying unique sequences within contigs of the *T. pallidum* genome, such as by using well-known computer analysis software, and by generating and testing probes or amplification primers consisting of the DF sequence in an appropriate diagnostic format which determines amplification or hybridization selectivity.

The sequences falling within the scope of the present invention are not limited to the specific sequences herein described, but also include allelic and species variations thereof. Allelic and species variations can be routinely determined by comparing the polynucleotide sequences provided in SEQ ID NOS:1-744, ORF IDs and ORFs within, a representative fragment thereof, or a nucleotide sequence at least 99% and preferably 99.9% identical to said polynucleotide sequences, with a sequence from another isolate of the same species. Furthermore, to accommodate codon variability, the invention includes nucleic acid molecules coding for the same amino acid sequences as do the specific ORFs disclosed herein. In other words, in the coding region of an ORF, substitution of one codon for another which encodes the same amino acid is expressly contemplated.

Any specific sequence disclosed herein can be readily screened for errors by resequencing a particular fragment, such as an ORF, in both directions (*i.e.*, sequence both strands). Alternatively, error screening can be performed by sequencing corresponding polynucleotides of *T. pallidum* origin isolated by using part or all of the fragments in question as a probe or primer.

Each of the ORFs of the *T. pallidum* genome within the ORF IDs of Tables 1, 2 and 3, and the EMFs found 5' to the ORFs, can be used as polynucleotide reagents in numerous ways. For example, the sequences can be used as diagnostic probes or diagnostic amplification primers to detect the presence of a specific microbe in a sample, particularly *T. pallidum*. Especially preferred in this regard are ORFs such as those of Table 3, which do not match previously characterized sequences from other organisms and thus are most likely to be highly selective for *T. pallidum*. Also particularly preferred are ORFs that can be used to distinguish between strains of *T. pallidum*, particularly those that distinguish medically important strain, such as drug-resistant strains.

In addition, the fragments of the present invention, as broadly described, can be used to control gene expression through triple helix formation or antisense DNA or RNA, both of which methods are based on the binding of a polynucleotide sequence to DNA or RNA. Triple helix-formation optimally results in a shut-off of RNA transcription from DNA, while antisense RNA hybridization blocks translation of an mRNA molecule into polypeptide. Information from the sequences of the present invention can be used to design antisense and triple helix-forming oligonucleotides. Polynucleotides suitable for use in these methods are usually 20 to 40 bases in

length and are designed to be complementary to a region of the gene involved in transcription, for triple-helix formation, or to the mRNA itself, for antisense inhibition. Both techniques have been demonstrated to be effective in model systems, and the requisite techniques are well known and involve routine procedures. Triple helix techniques are discussed in, for example, Lee *et al.*, *Nucl. Acids Res. 6*:3073 (1979); Cooney *et al.*, *Science 241*:456 (1988); and Dervan *et al.*, *Science 251*:1360 (1991). Antisense techniques in general are discussed in, for instance, Okano, *J. Neurochem. 56*:560 (1991) and *Oligodeoxynucleotides as Antisense Inhibitors of Gene Expression*, CRC Press, Boca Raton, FL (1988)).

The present invention further provides recombinant constructs comprising one or more fragments of the *T. pallidum* genomic fragments and contigs of the present invention. Certain preferred recombinant constructs of the present invention comprise a vector, such as a plasmid or viral vector, into which a fragment of the *T. pallidum* genome has been inserted, in a forward or reverse orientation. In the case of a vector comprising one of the ORFs of the present invention, the vector may further comprise regulatory sequences, including for example, a promoter, operably linked to the ORF. For vectors comprising the EMFs of the present invention, the vector may further comprise a marker sequence or heterologous ORF operably linked to the EMF.

Large numbers of suitable vectors and promoters are known to those of skill in the art and are commercially available for generating the recombinant constructs of the present invention. The following vectors are provided by way of example. Useful bacterial vectors include phagescript, PsiX174, pBluescript SK, pBS KS, pNH8a, pNH16a, pNH18a, pNH46a (available from Stratagene); pTrc99A, pKK223-3, pKK233-3, pDR540, pRIT5 (available from Pharmacia). Useful eukaryotic vectors include pWLneo, pSV2cat, pOG44, pXT1, pSG (available from Stratagene) pSVK3, pBPV, pMSG, pSVL (available from Pharmacia).

Promoter regions can be selected from any desired gene using CAT (chloramphenicol transferase) vectors or other vectors with selectable markers. Two appropriate vectors are pKK232-8 and pCM7. Particular named bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda PR, and trc. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein- I. Selection of the appropriate vector and promoter is well within the level of ordinary skill in the art.

The present invention further provides host cells containing any one of the isolated fragments of the *T. pallidum* genomic fragments and contigs of the present invention, wherein the fragment has been introduced into the host cell using known methods. The host cell can be a higher eukaryotic host cell, such as a mammalian cell, a lower eukaryotic host cell, such as a yeast cell, or a procaryotic cell, such as a bacterial cell.

A polynucleotide of the present invention, such as a recombinant construct comprising an ORF of the present invention, may be introduced into the host by a variety of well established techniques that are standard in the art, such as calcium phosphate transfection, DEAE, dextran

mediated transfection and electroporation, which are described in, for instance, Davis, L. *et al.*, BASIC METHODS IN MOLECULAR BIOLOGY (1986).

A host cell containing one of the fragments of the *T. pallidum* genomic fragments and contigs of the present invention, can be used in conventional manners to produce the gene product encoded by the isolated fragment (in the case of an ORF) or can be used to produce a heterologous protein under the control of the EMF.

The present invention further provides isolated polypeptides encoded by the nucleic acid fragments of the present invention or by degenerate variants of the nucleic acid fragments of the present invention. By "degenerate variant" is intended nucleotide fragments which differ from a nucleic acid fragment of the present invention (*e.g.*, an ORF) by nucleotide sequence but, due to the degeneracy of the Genetic Code, encode an identical polypeptide sequence.

Preferred nucleic acid fragments of the present invention are the ORF IDs depicted in Tables 2 and 3 and the ORFs within which encode proteins.

A variety of methodologies known in the art can be utilized to obtain any one of the isolated polypeptides or proteins of the present invention. At the simplest level, the amino acid sequence can be synthesized using commercially available peptide synthesizers. This is particularly useful in producing small peptides and fragments of larger polypeptides. Such short fragments as may be obtained most readily by synthesis are useful, for example, in generating antibodies against the native polypeptide, as discussed further below.

In an alternative method, the polypeptide or protein is purified from bacterial cells which naturally produce the polypeptide or protein. One skilled in the art can readily employ well-known methods for isolating polypeptides and proteins to isolate and purify polypeptides or proteins of the present invention produced naturally by a bacterial strain, or by other methods. Methods for isolation and purification that can be employed in this regard include, but are not limited to, immunochromatography, HPLC, size-exclusion chromatography, ion-exchange chromatography, and immuno-affinity chromatography.

The polypeptides and proteins of the present invention also can be purified from cells which have been altered to express the desired polypeptide or protein. As used herein, a cell is said to be altered to express a desired polypeptide or protein when the cell, through genetic manipulation, is made to produce a polypeptide or protein which it normally does not produce or which the cell normally produces at a lower level. Those skilled in the art can readily adapt procedures for introducing and expressing either recombinant or synthetic sequences into eukaryotic or prokaryotic cells in order to generate a cell which produces one of the polypeptides or proteins of the present invention.

The polypeptides of the present invention are preferably provided in an isolated form, and preferably are substantially purified. A recombinantly produced version of the *T. pallidum* polypeptide can be substantially purified by the one-step method described by Smith et al. (1988) Gene 67:31-40. Polypeptides of the invention also can be purified from natural or recombinant sources using antibodies directed against the polypeptides of the invention in methods which are

well known in the art of protein purification.

The invention further provides for isolated *T. pallidum* polypeptides comprising an amino acid sequence selected from the group including: (a) the amino acid sequence of a full-length *T. pallidum* polypeptide having the complete amino acid sequence from the first methionine codon to the termination codon of each sequence listed in SEQ ID NOS:1-744, wherein said termination codon is at the end of each SEQ ID NO: and said first methionine is the first methionine in frame with said termination codon; and (b) the amino acid sequence of a full-length *T. pallidum* polypeptide having the complete amino acid sequence in (a) excepting the N-terminal methionine.

The polypeptides of the present invention also include polypeptides having an amino acid sequence at least 80% identical, more preferably at least 90% identical, and still more preferably 95%, 96%, 97%, 98% or 99% identical to those described in (a) and (b) above.

The present invention is further directed to polynucleotides encoding portions or fragments of the amino acid sequences described herein as well as to portions or fragments of the isolated amino acid sequences described herein. Fragments include portions of the amino acid sequences described herein at least 5 contiguous amino acid in length and selected from any two integers, one of which representing an N-terminal position and another representing a C-terminal position. The initiation codon of the ORFs of the present invention is position 1. The initiation codon (positon 1) for purposes of the present invention is the first methionine codon of each ORF ID which is in frame with the termination codon at the end of each said sequence. Every combination of a N-terminal and C-terminal position that a fragment at least 5 contiguous amino acid residues in length could occupy, on any given ORF is included in the invention, i.e., from initiation codon up to the termination codon. "At least" means a fragment may be 5 contiguous amino acid residues in length or any integer between 5 and the number of residues in an ORF, minus 1. Therefore, included in the invention are contiguous fragments specified by any N-terminal and C-terminal positions of amino acid sequence set forth in SEQ ID NOS:1-744 or Tables 1-3 wherein the contiguous fragment is any integer between 5 and the number of residues in an ORF minus 1.

Further, the invention includes polypeptides comprising fragments specified by size, in amino acid residues, rather than by N-terminal and C-terminal positions. The invention includes any fragment size, in contiguous amino acid residues, selected from integers between 5 and the number of residues in an ORF, minus 1. Preferred sizes of contiguous polypeptide fragments include about 5 amino acid residues, about 10 amino acid residues, about 20 amino acid residues, about 30 amino acid residues, about 40 amino acid residues, about 50 amino acid residues, about 100 amino acid residues, about 200 amino acid residues, about 300 amino acid residues, and about 400 amino acid residues. The preferred sizes are, of course, meant to exemplify, not limit, the present invention as all size fragments representing any integer between 5 and the number of residues in a full length sequence minus 1 are included in the invention. The present invention also provides for the exclusion of any fragments specified by N-terminal and C-terminal positions or by size in amino acid residues as described above. Any number of fragments

specified by N-terminal and C-terminal positions or by size in amino acid residues as described above may be excluded.

The above fragments need not be active since they would be useful, for example, in immunoassays, in epitope mapping, epitope tagging, to generate antibodies to a particular portion of the protein, as vaccines, and as molecular weight markers.

Further polypeptides of the present invention include polypeptides which have at least 90% similarity, more preferably at least 95% similarity, and still more preferably at least 96%, 97%, 98% or 99% similarity to those described above.

A further embodiment of the invention relates to a polypeptide which comprises the amino acid sequence of a *T. pallidum* polypeptide having an amino acid sequence which contains at least one conservative amino acid substitution, but not more than 50 conservative amino acid substitutions, not more than 40 conservative amino acid substitutions, not more than 30 conservative amino acid substitutions, and not more than 20 conservative amino acid substitutions. Also provided are polypeptides which comprise the amino acid sequence of a *T. pallidum* polypeptide, having at least one, but not more than 10, 9, 8, 7, 6, 5, 4, 3, 2 or 1 conservative amino acid substitutions.

By a polypeptide having an amino acid sequence at least, for example, 95% "identical" to a query amino acid sequence of the present invention, it is intended that the amino acid sequence of the subject polypeptide is identical to the query sequence except that the subject polypeptide sequence may include up to five amino acid alterations per each 100 amino acids of the query amino acid sequence. In other words, to obtain a polypeptide having an amino acid sequence at least 95% identical to a query amino acid sequence, up to 5% of the amino acid residues in the subject sequence may be inserted, deleted, (indels) or substituted with another amino acid. These alterations of the reference sequence may occur at the amino or carboxy terminal positions of the reference amino acid sequence or anywhere between those terminal positions, interspersed either individually among residues in the reference sequence.

As a practical matter, whether any particular polypeptide is at least 90%, 95%, 96%, 97%, 98% or 99% identical to the ORF amino acid sequences encoded by the sequences of SEQ ID NOS:1-744, as described hererin, can be determined conventionally using known computer programs. A preferred method for determining the best overall match between a query sequence (a sequence of the present invention) and a subject sequence, also referred to as a global sequence alignment, can be determined using the FASTDB computer program based on the algorithm of Brutlag et al., (1990) Comp. App. Biosci. 6:237-245. In a sequence alignment the query and subject sequences are both amino acid sequences. The result of said global sequence alignment is in percent identity. Preferred parameters used in a FASTDB amino acid alignment are: Matrix=PAM 0, k-tuple=2, Mismatch Penalty=1, Joining Penalty=20, Randomization Group Length=0, Cutoff Score=1, Window Size=sequence length, Gap Penalty=5, Gap Size Penalty=0.05, Window Size=500 or the length of the subject amino acid sequence, whichever is shorter.

If the subject sequence is shorter than the query sequence due to N- or C-terminal deletions, not because of internal deletions, the results, in percent identity, must be manually corrected. This is because the FASTDB program does not account for N- and C-terminal truncations of the subject sequence when calculating global percent identity. For subject sequences truncated at the N- and C-termini, relative to the query sequence, the percent identity is corrected by calculating the number of residues of the query sequence that are N- and C-terminal of the subject sequence, which are not matched/aligned with a corresponding subject residue, as a percent of the total bases of the query sequence. Whether a residue is matched/aligned is determined by results of the FASTDB sequence alignment. This percentage is then subtracted from the percent identity, calculated by the above FASTDB program using the specified parameters, to arrive at a final percent identity score. This final percent identity score is what is used for the purposes of the present invention. Only residues to the N- and C-termini of the subject sequence, which are not matched/aligned with the query sequence, are considered for the purposes of manually adjusting the percent identity score. That is, only query amino acid residues outside the farthest N- and C-terminal residues of the subject sequence.

For example, a 90 amino acid residue subject sequence is aligned with a 100 residue query sequence to determine percent identity. The deletion occurs at the N-terminus of the subject sequence and therefore, the FASTDB alignment does not match/align with the first 10 residues at the N-terminus. The 10 unpaired residues represent 10% of the sequence (number of residues at the N- and C- termini not matched/total number of residues in the query sequence) so 10% is subtracted from the percent identity score calculated by the FASTDB program. If the remaining 90 residues were perfectly matched the final percent identity would be 90%. In another example, a 90 residue subject sequence is compared with a 100 residue query sequence. This time the deletions are internal so there are no residues at the N- or C-termini of the subject sequence which are not matched/aligned with the query. In this case the percent identity calculated by FASTDB is not manually corrected. Once again, only residue positions outside the N- and C-terminal ends of the subject sequence, as displayed in the FASTDB alignment, which are not matched/aligned with the query sequence are manually corrected. No other manual corrections are to made for the purposes of the present invention.

The above polypeptide sequences are included irrespective of whether they have their normal biological activity. This is because even where a particular polypeptide molecule does not have biological activity, one of skill in the art would still know how to use the polypeptide, for instance, as a vaccine or to generate antibodies. Other uses of the polypeptides of the present invention that do not have *T. pallidum* activity include, *inter alia*, as epitope tags, in epitope mapping, and as molecular weight markers on SDS-PAGE gels or on molecular sieve gel filtration columns using methods known to those of skill in the art.

As described below, the polypeptides of the present invention can also be used to raise polyclonal and monoclonal antibodies, which are useful in assays for detecting *T. pallidum* protein expression or as agonists and antagonists capable of enhancing or inhibiting *T. pallidum* protein

function. Further, such polypeptides can be used in the yeast two-hybrid system to "capture" *T. pallidum* protein binding proteins which are also candidate agonists and antagonists according to the present invention. *See, e.g.,* Fields et al. (1989) Nature 340:245-246.

Any host/vector system can be used to express one or more of the ORFs of the present invention. These include, but are not limited to, eukaryotic hosts such as HeLa cells, CV-1 cell, COS cells, and Sf9 cells, as well as prokaryotic host such as *E. coli* and *B. subtilis*. The most preferred cells are those which do not normally express the particular polypeptide or protein or which expresses the polypeptide or protein at low natural level.

"Recombinant," as used herein, means that a polypeptide or protein is derived from recombinant (*e.g.*, microbial or mammalian) expression systems. "Microbial" refers to recombinant polypeptides or proteins made in bacterial or fungal (*e.g.*, yeast) expression systems. As a product, "recombinant microbial"defines a polypeptide or protein essentially free of native endogenous substances and unaccompanied by associated native glycosylation. Polypeptides or proteins expressed in most bacterial cultures, *e.g.*, *E. coli*, will be free of glycosylation modifications; polypeptides or proteins expressed in yeast will have a glycosylation pattern different from that expressed in mammalian cells.

"Nucleotide sequence" refers to a heteropolymer of deoxyribonucleotides. Generally, DNA segments encoding the polypeptides and proteins provided by this invention are assembled from fragments of the *T. pallidum* genome and short oligonucleotide linkers, or from a series of oligonucleotides, to provide a synthetic gene which is capable of being expressed in a recombinant transcriptional unit comprising regulatory elements derived from a microbial or viral operon.

Recombinant expression vehicle or vector" refers to a plasmid or phage or virus or vector, for expressing a polypeptide from a DNA (RNA) sequence. The expression vehicle can comprise a transcriptional unit comprising an assembly of (1) a genetic regulatory elements necessary for gene expression in the host, including elements required to initiate and maintain transcription at a level sufficient for suitable expression of the desired polypeptide, including, for example, promoters and, where necessary, an enhancer and a polyadenylation signal; (2) a structural or coding sequence which is transcribed into mRNA and translated into protein, and (3) appropriate signals to initiate translation at the beginning of the desired coding region and terminate translation at its end. Structural units intended for use in yeast or eukaryotic expression systems preferably include a leader sequence enabling extracellular secretion of translated protein by a host cell. Alternatively, where recombinant protein is expressed without a leader or transport sequence, it may include an N-terminal methionine residue. This residue may or may not be subsequently cleaved from the expressed recombinant protein to provide a final product.

"Recombinant expression system" means host cells which have stably integrated a recombinant transcriptional unit into chromosomal DNA or carry the recombinant transcriptional unit extra chromosomally. The cells can be prokaryotic or eukaryotic. Recombinant expression

systems as defined herein will express heterologous polypeptides or proteins upon induction of the regulatory elements linked to the DNA segment or synthetic gene to be expressed.

Mature proteins can be expressed in mammalian cells, yeast, bacteria, or other cells under the control of appropriate promoters. Cell-free translation systems can also be employed to produce such proteins using RNAs derived from the DNA constructs of the present invention. Appropriate cloning and expression vectors for use with prokaryotic and eukaryotic hosts are described in Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual*, 2nd Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (1989), the disclosure of which is hereby incorporated by reference in its entirety.

Generally, recombinant expression vectors will include origins of replication and selectable markers permitting transformation of the host cell, *e.g.*, the ampicillin resistance gene of *E. coli* and *S. cerevisiae* TRP1 gene, and a promoter derived from a highly expressed gene to direct transcription of a downstream structural sequence. Such promoters can be derived from operons encoding glycolytic enzymes such as 3- phosphoglycerate kinase (PGK), alpha-factor, acid phosphatase, or heat shock proteins, among others. The heterologous structural sequence is assembled in appropriate phase with translation initiation and termination sequences, and preferably, a leader sequence capable of directing secretion of translated protein into the periplasmic space or extracellular medium. Optionally, the heterologous sequence can encode a fusion protein including an N-terminal identification peptide imparting desired characteristics, e.g., stabilization or simplified purification of expressed recombinant product.

Useful expression vectors for bacterial use are constructed by inserting a structural DNA sequence encoding a desired protein together with suitable translation initiation and termination signals in operable reading phase with a functional promoter. The vector will comprise one or more phenotypic selectable markers and an origin of replication to ensure maintenance of the vector and, when desirable, provide amplification within the host.

Suitable prokaryotic hosts for transformation include strains of *E. coli*, *B. subtilis*, *Salmonella typhimurium* and various species within the genera *Pseudomonas* and *Streptomyces*. Others may, also be employed as a matter of choice.

As a representative but non-limiting example, useful expression vectors for bacterial use can comprise a selectable marker and bacterial origin of replication derived from commercially available plasmids comprising genetic elements of the well known cloning vector pBR322 (ATCC 37017). Such commercial vectors include, for example, pKK223-3 (available form Pharmacia Fine Chemicals, Uppsala, Sweden) and GEM 1 (available from Promega Biotec, Madison, WI, USA). These pBR322 "backbone" sections are combined with an appropriate promoter and the structural sequence to be expressed.

Following transformation of a suitable host strain and growth of the host strain to an appropriate cell density, the selected promoter, where it is inducible, is derepressed or induced by appropriate means (*e.g.*, temperature shift or chemical induction) and cells are cultured for an additional period to provide for expression of the induced gene product. Thereafter cells are

typically harvested, generally by centrifugation, disrupted to release expressed protein, generally by physical or chemical means, and the resulting crude extract is retained for further purification.

Various mammalian cell culture systems can also be employed to express recombinant protein. Examples of mammalian expression systems include the COS-7 lines of monkey kidney fibroblasts, described in Gluzman, *Cell 23:*175 (1981), and other cell lines capable of expressing a compatible vector, for example, the C127, 3T3, CHO, HeLa and BHK cell lines.

Mammalian expression vectors will comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5 flanking nontranscribed sequences. DNA sequences derived from the SV40 viral genome, for example, SV40 origin, early promoter, enhancer, splice, and polyadenylation sites may be used to provide the required nontranscribed genetic elements.

Recombinant polypeptides and proteins produced in bacterial culture is usually isolated by initial extraction from cell pellets, followed by one or more salting-out, aqueous ion exchange or size exclusion chromatography steps. Microbial cells employed in expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents. Protein refolding steps can be used, as necessary, in completing configuration of the mature protein. Finally, high performance liquid chromatography (HPLC) can be employed for final purification steps.

The present invention further includes isolated polypeptides, proteins and nucleic acid molecules which are substantially equivalent to those herein described. As used herein, substantially equivalent can refer both to nucleic acid and amino acid sequences, for example a mutant sequence, that varies from a reference sequence by one or more substitutions, deletions, or additions, the net effect of which does not result in an adverse functional dissimilarity between reference and subject sequences. For purposes of the present invention, sequences having equivalent biological activity, and equivalent expression characteristics are considered substantially equivalent. For purposes of determining equivalence, truncation of the mature sequence should be disregarded.

The invention further provides methods of obtaining homologs from other strains of *T. pallidum*, of the fragments of the *T. pallidum* genome of the present invention and homologs of the proteins encoded by the ORFs of the present invention. As used herein, a sequence or protein of *T. pallidum* is defined as a homolog of a fragment of the *T. pallidum* fragments or contigs or a protein encoded by one of the ORFs of the present invention, if it shares significant homology to one of the fragments of the *T. pallidum* genome of the present invention or a protein encoded by one of the ORFs of the present invention. Specifically, by using the sequence disclosed herein as a probe or as primers, and techniques such as PCR cloning and colony/plaque hybridization, one skilled in the art can obtain homologs.

As used herein, two nucleic acid molecules or proteins are said to "share significant homology" if the two contain regions which possess greater than 85% sequence (amino acid or

nucleic acid) homology. Preferred homologs in this regard are those with more than 90% homology. Especially preferred are those with 93% or more homology. Among especially preferred homologs those with 95% or more homology are particularly preferred. Very particularly preferred among these are those with 97% and even more particularly preferred among those are homologs with 99% or more homology. The most preferred homologs among these are those with 99.9% homology or more. It will be understood that, among measures of homology, identity is particularly preferred in this regard.

Region specific primers or probes derived from the nucleotide sequence provided in SEQ ID NOS: 1-744 or from a nucleotide sequence at least 95%, particularly at least 99%, especially at least 99.5% identical to a sequence of SEQ ID NOS: 1-744 can be used to prime DNA synthesis and PCR amplification, as well as to identify colonies containing cloned DNA encoding a homolog. Methods suitable to this aspect of the present invention are well known and have been described in great detail in many publications such as, for example, Innis *et al.*, *PCR Protocols*, Academic Press, San Diego, CA (1990)).

When using primers derived from SEQ ID NOS: 1-744 or from a nucleotide sequence having an aforementioned identity to a sequence of SEQ ID NOS:1-744, one skilled in the art will recognize that by employing high stringency conditions (*e.g.*, annealing at 50-60°C in 6X SSPC and 50% formamide, and washing at 50- 65°C in 0.5X SSPC) only sequences which are greater than 75% homologous to the primer will be amplified. By employing lower stringency conditions (*e.g.*, hybridizing at 35-37°C in 5X SSPC and 40-45% formamide, and washing at 42°C in 0.5X SSPC), sequences which are greater than 40-50% homologous to the primer will also be amplified.

When using DNA probes derived from SEQ ID NOS:1-744, or from a nucleotide sequence having an aforementioned identity to a sequence of SEQ ID NOS: 1-744 , for colony/plaque hybridization, one skilled in the art will recognize that by employing high stringency conditions (*e.g.*, hybridizing at 50- 65°C in 5X SSPC and 50% formamide, and washing at 50- 65°C in 0.5X SSPC), sequences having regions which are greater than 90% homologous to the probe can be obtained, and that by employing lower stringency conditions (*e.g.*, hybridizing at 35-37°C in 5X SSPC and 40-45% formamide, and washing at 42°C in 0.5X SSPC), sequences having regions which are greater than 35-45% homologous to the probe will be obtained.

Any organism can be used as the source for homologs of the present invention so long as the organism naturally expresses such a protein or contains genes encoding the same. The most preferred organism for isolating homologs are bacteria which are closely related to *T. pallidum*.

ILLUSTRATIVE USES OF COMPOSITIONS OF THE INVENTION

Each ORF corresponding to the ORF IDs provided in Tables 1 and 2 is identified with a function by homology to a known gene or polypeptide. As a result, one skilled in the art can use

the polypeptides of the present invention for commercial, therapeutic and industrial purposes consistent with the type of putative identification of the polypeptide. Such identifications permit one skilled in the art to use the *T. pallidum* ORFs in a manner similar to the known type of sequences for which the identification is made; for example, to ferment a particular sugar source or to produce a particular metabolite. A variety of reviews illustrative of this aspect of the invention are available, including the following reviews on the industrial use of enzymes, for example, BIOCHEMICAL ENGINEERING AND BIOTECHNOLOGY HANDBOOK, 2nd Ed., MacMillan Publications, Ltd. NY (1991) and BIOCATALYSTS IN ORGANIC SYNTHESES, Tramper *et al.*, Eds., Elsevier Science Publishers, Amsterdam, The Netherlands (1985). A variety of exemplary uses that illustrate this and similar aspects of the present invention are discussed below.

1. Biosynthetic Enzymes

Open reading frames encoding proteins involved in mediating the catalytic reactions involved in intermediary and macromolecular metabolism, the biosynthesis of small molecules, cellular processes and other functions includes enzymes involved in the degradation of the intermediary products of metabolism, enzymes involved in central intermediary metabolism, enzymes involved in respiration, both aerobic and anaerobic, enzymes involved in fermentation, enzymes involved in ATP proton motor force conversion, enzymes involved in broad regulatory function, enzymes involved in amino acid synthesis, enzymes involved in nucleotide synthesis, enzymes involved in cofactor and vitamin synthesis, can be used for industrial biosynthesis.

The various metabolic pathways present in *T. pallidum* can be identified based on absolute nutritional requirements as well as by examining the various enzymes identified in Table 1-3 and SEQ ID NOS:1-744.

Of particular interest are polypeptides involved in the degradation of intermediary metabolites as well as non-macromolecular metabolism. Such enzymes include amylases, glucose oxidases, and catalase.

Proteolytic enzymes are another class of commercially important enzymes. Proteolytic enzymes find use in a number of industrial processes including the processing of flax and other vegetable fibers, in the extraction, clarification and depectinization of fruit juices, in the extraction of vegetables' oil and in the maceration of fruits and vegetables to give unicellular fruits. A detailed review of the proteolytic enzymes used in the food industry is provided in Rombouts *et al.*, *Symbiosis 21:*79 (1986) and Voragen *et al.* in *Biocatalysts In Agricultural Biotechnology*, Whitaker *et al.*, Eds., *American Chemical Society Symposium Series 389:*93 (1989) .

The metabolism of sugars is an important aspect of the primary metabolism of *T. pallidum.* Enzymes involved in the degradation of sugars, such as, particularly, glucose, galactose, fructose and xylose, can be used in industrial fermentation. Some of the important sugar transforming enzymes, from a commercial viewpoint, include sugar isomerases such as glucose isomerase. Other metabolic enzymes have found commercial use such as glucose

oxidases which produces ketogulonic acid (KGA). KGA is an intermediate in the commercial production of ascorbic acid using the Reichstein's procedure, as described in Krueger *et al.*, *Biotechnology* 6(A), Rhine *et al.*, Eds., Verlag Press, Weinheim, Germany (1984).

Glucose oxidase (GOD) is commercially available and has been used in purified form as well as in an immobilized form for the deoxygenation of beer. See, for instance, Hartmeir *et al.*, *Biotechnology Letters 1*:21 (1979). The most important application of GOD is the industrial scale fermentation of gluconic acid. Market for gluconic acids which are used in the detergent, textile, leather, photographic, pharmaceutical, food, feed and concrete industry, as described, for example, in Bigelis *et al.*, beginning on page 357 in GENE MANIPULATIONS AND FUNGI; Benett *et al.*, Eds., Academic Press, New York (1985). In addition to industrial applications, GOD has found applications in medicine for quantitative determination of glucose in body fluids recently in biotechnology for analyzing syrups from starch and cellulose hydrosylates. This application is described in Owusu *et al.*, *Biochem. et Biophysica. Acta. 872:*83 (1986), for instance.

The main sweetener used in the world today is sugar which comes from sugar beets and sugar cane. In the field of industrial enzymes, the glucose isomerase process shows the largest expansion in the market today. Initially, soluble enzymes were used and later immobilized enzymes were developed (Krueger *et al.*, *Biotechnology, The Textbook of Industrial Microbiology*, Sinauer Associated Incorporated, Sunderland, Massachusetts (1990)). Today, the use of glucose- produced high fructose syrups is by far the largest industrial business using immobilized enzymes. A review of the industrial use of these enzymes is provided by Jorgensen, Starch *40*:307 (1988).

Proteinases, such as alkaline serine proteinases, are used as detergent additives and thus represent one of the largest volumes of microbial enzymes used in the industrial sector. Because of their industrial importance, there is a large body of published and unpublished information regarding the use of these enzymes in industrial processes. (See Faultman *et al.*, Acid Proteases Structure Function and Biology, Tang, J., ed., Plenum Press, New York (1977) and Godfrey *et al.*, Industrial Enzymes, MacMillan Publishers, Surrey, UK (1983) and Hepner *et al.*, Report Industrial Enzymes by 1990, Hel Hepner & Associates, London (1986)).

Another class of commercially usable proteins of the present invention are the microbial lipases, described by, for instance, Macrae *et al.*, *Philosophical Transactions of the Chiral Society of London 310*:227 (1985) and Poserke, *Journal of the American Oil Chemist Society 61:*1758 (1984). A major use of lipases is in the fat and oil industry for the production of neutral glycerides using lipase catalyzed inter-esterification of readily available triglycerides. Application of lipases include the use as a detergent additive to facilitate the removal of fats from fabrics in the course of the washing procedures.

The use of enzymes, and in particular microbial enzymes, as catalyst for key steps in the synthesis of complex organic molecules is gaining popularity at a great rate. One area of great interest is the preparation of chiral intermediates. Preparation of chiral intermediates is of interest

to a wide range of synthetic chemists particularly those scientists involved with the preparation of new pharmaceuticals, agrochemicals, fragrances and flavors. (See Davies *et al.*, *Recent Advances in the Generation of Chiral Intermediates Using Enzymes*, CRC Press, Boca Raton, Florida (1990)). The following reactions catalyzed by enzymes are of interest to organic chemists: hydrolysis of carboxylic acid esters, phosphate esters, amides and nitriles, esterification reactions, trans-esterification reactions, synthesis of amides, reduction of alkanones and oxoalkanates, oxidation of alcohols to carbonyl compounds, oxidation of sulfides to sulfoxides, and carbon bond forming reactions such as the aldol reaction.

When considering the use of an enzyme encoded by one of the ORFs of the present invention for biotransformation and organic synthesis it is sometimes necessary to consider the respective advantages and disadvantages of using a microorganism as opposed to an isolated enzyme. Pros and cons of using a whole cell system on the one hand or an isolated partially purified enzyme on the other hand, has been described in detail by Bud *et al.*, Chemistry in Britain (1987), p. 127.

Amino transferases, enzymes involved in the biosynthesis and metabolism of amino acids, are useful in the catalytic production of amino acids. The advantages of using microbial based enzyme systems is that the amino transferase enzymes catalyze the stereo- selective synthesis of only L-amino acids and generally possess uniformly high catalytic rates. A description of the use of amino transferases for amino acid production is provided by Roselle-David, *Methods of Enzymology 136*:479 (1987).

Another category of useful proteins encoded by the ORFs of the present invention include enzymes involved in nucleic acid synthesis, repair, and recombination.

2. Generation of Antibodies

As described here, the proteins of the present invention, as well as homologs thereof, can be used in a variety of procedures and methods known in the art which are currently applied to other proteins. The proteins of the present invention can further be used to generate an antibody which selectively binds the protein.

T. pallidum protein-specific antibodies for use in the present invention can be raised against the intact *T. pallidum* protein or an antigenic polypeptide fragment thereof, which may be presented together with a carrier protein, such as an albumin, to an animal system (such as rabbit or mouse) or, if it is long enough (at least about 25 amino acids), without a carrier.

As used herein, the term "antibody" (Ab) or "monoclonal antibody" (Mab) is meant to include intact molecules, single chain whole antibodies, and antibody fragments. Antibody fragments of the present invention include Fab and F(ab')2 and other fragments including single-chain Fvs (scFv) and disulfide-linked Fvs (sdFv). Also included in the present invention are chimeric and humanized monoclonal antibodies and polyclonal antibodies specific for the polypeptides of the present invention. The antibodies of the present invention may be prepared by any of a variety of methods. For example, cells expressing a polypeptide of the present

invention or an antigenic fragment thereof can be administered to an animal in order to induce the production of sera containing polyclonal antibodies. For example, a preparation of *T. pallidum* polypeptide or fragment thereof is prepared and purified to render it substantially free of natural contaminants. Such a preparation is then introduced into an animal in order to produce polyclonal antisera of greater specific activity.

In a preferred method, the antibodies of the present invention are monoclonal antibodies or binding fragments thereof. Such monoclonal antibodies can be prepared using hybridoma technology. *See, e.g.*, Harlow et al., ANTIBODIES: A LABORATORY MANUAL, (Cold Spring Harbor Laboratory Press, 2nd ed. 1988); Hammerling, et al., in: MONOCLONAL ANTIBODIES AND T-CELL HYBRIDOMAS 563-681 (Elsevier, N.Y., 1981). Fab and F(ab')2 fragments may be produced by proteolytic cleavage, using enzymes such as papain (to produce Fab fragments) or pepsin (to produce F(ab')2 fragments). Alternatively, *T. pallidum* polypeptide-binding fragments, chimeric, and humanized antibodies can be produced through the application of recombinant DNA technology or through synthetic chemistry using methods known in the art.

Alternatively, additional antibodies capable of binding to the polypeptide antigen of the present invention may be produced in a two-step procedure through the use of anti-idiotypic antibodies. Such a method makes use of the fact that antibodies are themselves antigens, and that, therefore, it is possible to obtain an antibody which binds to a second antibody. In accordance with this method, *T. pallidum* polypeptide-specific antibodies are used to immunize an animal, preferably a mouse. The splenocytes of such an animal are then used to produce hybridoma cells, and the hybridoma cells are screened to identify clones which produce an antibody whose ability to bind to the *T. pallidum* polypeptide-specific antibody can be blocked by the *T. pallidum* polypeptide antigen. Such antibodies comprise anti-idiotypic antibodies to the *T. pallidum* polypeptide-specific antibody and can be used to immunize an animal to induce formation of further *T. pallidum* polypeptide-specific antibodies.

Antibodies and fragements thereof of the present invention may be described by the portion of a polypeptide of the present invention recognized or specifically bound by the antibody. Antibody binding fragements of a polypeptide of the present invention may be described or specified in the same manner as for polypeptide fragements discussed above., i.e, by N-terminal and C-terminal positions or by size in contiguous amino acid residues. Any number of antibody binding fragments, of a polypeptide of the present invention, specified by N-terminal and C-terminal positions or by size in amino acid residues, as described above, may also be excluded from the present invention. Therefore, the present invention includes antibodies the specifically bind a particuarlly discribed fragement of a polypeptide of the present invention and allows for the exclusion of the same.

Antibodies and fragements thereof of the present invention may also be described or specified in terms of their cross-reactivity. Antibodies and fragements that do not bind polypeptides of any other species of *Borrelia* other than *T. pallidum* are included in the present invention. Likewise,

antibodies and fragements that bind only species of *Borrelia*, i.e. antibodies and fragements that do not bind bacteria from any genus other than *Borrelia*, are included in the present invention.

The present invention further provides the above- described antibodies in detectably labelled form. Antibodies can be detectably labelled through the use of radioisotopes, affinity labels (such as biotin, avidin, *etc.*), enzymatic labels (such as horseradish peroxidase, alkaline phosphatase, *etc.*) fluorescent labels (such as FITC or rhodamine, *etc.*), paramagnetic atoms, *etc.* Procedures for accomplishing such labeling are well-known in the art, for example see Sternberger *et al.*, *J. Histochem. Cytochem. 18*:315 (1970); Bayer, E. A. *et al.*, *Meth. Enzym. 62*:308 (1979); Engval, E. *et al.*, *Immunol. 109*:129 (1972); Goding, J. W., *J. Immunol. Meth. 13*:215 (1976)).

The labeled antibodies of the present invention can be used for *in vitro*, *in vivo*, and in situ assays to identify cells or tissues in which a fragment of the *T. pallidum* genome is expressed.

The present invention further provides the above-described antibodies immobilized on a solid support. Examples of such solid supports include plastics such as polycarbonate, complex carbohydrates such as agarose and sepharose, acrylic resins and such as polyacrylamide and latex beads. Techniques for coupling antibodies to such solid supports are well known in the art (Weir, D. M. *et al.*, "Handbook of Experimental Immunology" 4th Ed., Blackwell Scientific Publications, Oxford, England, Chapter 10 (1986); Jacoby, W. D. *et al.*, Meth. Enzym. 34 Academic Press, N. Y. (1974)). The immobilized antibodies of the present invention can be used for *in vitro*, *in vivo*, and in situ assays as well as for immunoaffinity purification of the proteins of the present invention.

3. Epitope-Bearing Portions

In another aspect, the invention provides peptides and polypeptides comprising epitope-bearing portions of the *T. pallidum* polypeptides of the present invention. These epitopes are immunogenic or antigenic epitopes of the polypeptides of the present invention. An "immunogenic epitope" is defined as a part of a protein that elicits an antibody response when the whole protein or polypeptide is the immunogen. These immunogenic epitopes are believed to be confined to a few loci on the molecule. On the other hand, a region of a protein molecule to which an antibody can bind is defined as an "antigenic determinant" or "antigenic epitope." The number of immunogenic epitopes of a protein generally is less than the number of antigenic epitopes. *See, e.g.*, Geysen, et al. (1983) Proc. Natl. Acad. Sci. USA 81:3998- 4002. Amino acid residues comprising anigenic epitopes may be determined by algorithms such as the the Jameson-Wolf analysis or similar algorithms or by *in vivo* testing for an antigenic response using the methods described herein or those known in the art.

As to the selection of peptides or polypeptides bearing an antigenic epitope (*i.e.*, that contain a region of a protein molecule to which an antibody can bind), it is well known in that art that relatively short synthetic peptides that mimic part of a protein sequence are routinely capable

of eliciting an antiserum that reacts with the partially mimicked protein. *See, e.g.,* Sutcliffe, et al., (1983) Science 219:660-666. Peptides capable of eliciting protein-reactive sera are frequently represented in the primary sequence of a protein, can be characterized by a set of simple chemical rules, and are confined neither to immunodominant regions of intact proteins (*i.e.*, immunogenic epitopes) nor to the amino or carboxyl terminals. Peptides that are extremely hydrophobic and those of six or fewer residues generally are ineffective at inducing antibodies that bind to the mimicked protein; longer, peptides, especially those containing proline residues, usually are effective. *See,* Sutcliffe, et al., *supra*, p. 661. For instance, 18 of 20 peptides designed according to these guidelines, containing 8-39 residues covering 75% of the sequence of the influenza virus hemagglutinin HA1 polypeptide chain, induced antibodies that reacted with the HA1 protein or intact virus; and 12/12 peptides from the MuLV polymerase and 18/18 from the rabies glycoprotein induced antibodies that precipitated the respective proteins.

Antigenic epitope-bearing peptides and polypeptides of the invention are therefore useful to raise antibodies, including monoclonal antibodies, that bind specifically to a polypeptide of the invention. Thus, a high proportion of hybridomas obtained by fusion of spleen cells from donors immunized with an antigen epitope-bearing peptide generally secrete antibody reactive with the native protein. *See* Sutcliffe, et al., *supra*, p. 663. The antibodies raised by antigenic epitope-bearing peptides or polypeptides are useful to detect the mimicked protein, and antibodies to different peptides may be used for tracking the fate of various regions of a protein precursor which undergoes post-translational processing. The peptides and anti-peptide antibodies may be used in a variety of qualitative or quantitative assays for the mimicked protein, for instance in competition assays since it has been shown that even short peptides (*e.g.*, about 9 amino acids) can bind and displace the larger peptides in immunoprecipitation assays. *See, e.g.,* Wilson, et al., (1984) Cell 37:767-778. The anti-peptide antibodies of the invention also are useful for purification of the mimicked protein, for instance, by adsorption chromatography using methods known in the art.

Antigenic epitope-bearing peptides and polypeptides of the invention designed according to the above guidelines preferably contain a sequence of at least seven, more preferably at least nine and most preferably between about 10 to about 50 amino acids (i.e. any integer between 7 and 50) contained within the amino acid sequence of a polypeptide of the invention. However, peptides or polypeptides comprising a larger portion of an amino acid sequence of a polypeptide of the invention, containing about 50 to about 100 amino acids, or any length up to and including the entire amino acid sequence of a polypeptide of the invention, also are considered epitope-bearing peptides or polypeptides of the invention and also are useful for inducing antibodies that react with the mimicked protein. Preferably, the amino acid sequence of the epitope-bearing peptide is selected to provide substantial solubility in aqueous solvents (*i.e.*, the sequence includes relatively hydrophilic residues and highly hydrophobic sequences are preferably avoided); and sequences containing proline residues are particularly preferred.

The epitope-bearing peptides and polypeptides of the present invention may be produced by any conventional means for making peptides or polypeptides including recombinant means using nucleic acid molecules of the invention. For instance, an epitope-bearing amino acid sequence of the present invention may be fused to a larger polypeptide which acts as a carrier during recombinant production and purification, as well as during immunization to produce anti-peptide antibodies. Epitope-bearing peptides also may be synthesized using known methods of chemical synthesis. For instance, Houghten has described a simple method for synthesis of large numbers of peptides, such as 10-20 mg of 248 different 13 residue peptides representing single amino acid variants of a segment of the HA1 polypeptide which were prepared and characterized (by ELISA-type binding studies) in less than four weeks (Houghten, R. A. Proc. Natl. Acad. Sci. USA 82:5131-5135 (1985)). This "Simultaneous Multiple Peptide Synthesis (SMPS)" process is further described in U.S. Patent No. 4,631,211 to Houghten and coworkers (1986). In this procedure the individual resins for the solid-phase synthesis of various peptides are contained in separate solvent-permeable packets, enabling the optimal use of the many identical repetitive steps involved in solid-phase methods. A completely manual procedure allows 500-1000 or more syntheses to be conducted simultaneously (Houghten et al. (1985) Proc. Natl. Acad. Sci. 82:5131-5135 at 5134.

Epitope-bearing peptides and polypeptides of the invention are used to induce antibodies according to methods well known in the art. *See, e.g.*, Sutcliffe, et al., *supra*;; Wilson, et al., *supra*;; and Bittle, et al. (1985) J. Gen. Virol. 66:2347-2354. Generally, animals may be immunized with free peptide; however, anti-peptide antibody titer may be boosted by coupling of the peptide to a macromolecular carrier, such as keyhole limpet hemacyanin (KLH) or tetanus toxoid. For instance, peptides containing cysteine may be coupled to carrier using a linker such as m-maleimidobenzoyl-N-hydroxysuccinimide ester (MBS), while other peptides may be coupled to carrier using a more general linking agent such as glutaraldehyde. Animals such as rabbits, rats and mice are immunized with either free or carrier-coupled peptides, for instance, by intraperitoneal and/or intradermal injection of emulsions containing about 100 μg peptide or carrier protein and Freund's adjuvant. Several booster injections may be needed, for instance, at intervals of about two weeks, to provide a useful titer of anti-peptide antibody which can be detected, for example, by ELISA assay using free peptide adsorbed to a solid surface. The titer of anti-peptide antibodies in serum from an immunized animal may be increased by selection of anti-peptide antibodies, for instance, by adsorption to the peptide on a solid support and elution of the selected antibodies according to methods well known in the art.

Immunogenic epitope-bearing peptides of the invention, *i.e.*, those parts of a protein that elicit an antibody response when the whole protein is the immunogen, are identified according to methods known in the art. For instance, Geysen, *et al.*, *supra*, discloses a procedure for rapid concurrent synthesis on solid supports of hundreds of peptides of sufficient purity to react in an ELISA. Interaction of synthesized peptides with antibodies is then easily detected without removing them from the support. In this manner a peptide bearing an immunogenic epitope of a

desired protein may be identified routinely by one of ordinary skill in the art. For instance, the immunologically important epitope in the coat protein of foot-and-mouth disease virus was located by Geysen *et al. supra* with a resolution of seven amino acids by synthesis of an overlapping set of all 208 possible hexapeptides covering the entire 213 amino acid sequence of the protein. Then, a complete replacement set of peptides in which all 20 amino acids were substituted in turn at every position within the epitope were synthesized, and the particular amino acids conferring specificity for the reaction with antibody were determined. Thus, peptide analogs of the epitope-bearing peptides of the invention can be made routinely by this method. U.S. Patent No. 4,708,781 to Geysen (1987) further describes this method of identifying a peptide bearing an immunogenic epitope of a desired protein.

Further still, U.S. Patent No. 5,194,392, to Geysen (1990), describes a general method of detecting or determining the sequence of monomers (amino acids or other compounds) which is a topological equivalent of the epitope (*i.e.*, a "mimotope") which is complementary to a particular paratope (antigen binding site) of an antibody of interest. More generally, U.S. Patent No. 4,433,092, also to Geysen (1989), describes a method of detecting or determining a sequence of monomers which is a topographical equivalent of a ligand which is complementary to the ligand binding site of a particular receptor of interest. Similarly, U.S. Patent No. 5,480,971 to Houghten, R. A. *et al.* (1996) discloses linear C_1-C_7-alkyl peralkylated oligopeptides and sets and libraries of such peptides, as well as methods for using such oligopeptide sets and libraries for determining the sequence of a peralkylated oligopeptide that preferentially binds to an acceptor molecule of interest. Thus, non-peptide analogs of the epitope-bearing peptides of the invention also can be made routinely by these methods. The entire disclosure of each document cited in this section on "Polypeptides and Fragments" is hereby incorporated herein by reference.

As one of skill in the art will appreciate, the polypeptides of the present invention and the epitope-bearing fragments thereof described above can be combined with parts of the constant domain of immunoglobulins (IgG), resulting in chimeric polypeptides. These fusion proteins facilitate purification and show an increased half-life *in vivo*. This has been shown, *e.g.*, for chimeric proteins consisting of the first two domains of the human CD4-polypeptide and various domains of the constant regions of the heavy or light chains of mammalian immunoglobulins. (EPA 0,394,827; Traunecker et al. (1988) Nature 331:84-86. Fusion proteins that have a disulfide-linked dimeric structure due to the IgG part can also be more efficient in binding and neutralizing other molecules than a monomeric *T. pallidum* polypeptide or fragment thereof alone. *See* Fountoulakis et al. (1995) J. Biochem. 270:3958-3964. Nucleic acids encoding the above epitopes of *T. pallidum* polypeptides can also be recombined with a gene of interest as an epitope tag to aid in detection and purification of the expressed polypeptide.

3. Diagnostic Assays and Kits

The present invention further relates to methods for assaying Borrelia infection in an animal by detecting the expression of genes encoding Borrelia polypeptides of the present invention. The methods comprise analyzing tissue or body fluid from the animal for *Borrelia*-specific antibodies, nucleic acids, or proteins. Analysis of nucleic acid specific to *Borrelia* is assayed by PCR or hybridization techniques using nucleic acid sequences of the present invention as either hybridization probes or primers. *See, e.g.,* Sambrook et al. Molecular cloning: A Laboratory Manual (Cold Spring Harbor Laboratory Press, 2nd ed., 1989, page 54 reference); Eremeeva et al. (1994) J. Clin. Microbiol. 32:803-810 (describing differentiation among spotted fever group *Rickettsiae* species by analysis of restriction fragment length polymorphism of PCR-amplified DNA) and Chen et al. 1994 J. Clin. Microbiol. 32:589-595 (detecting *T. pallidum* nucleic acids *via* PCR).

Where diagnosis of a disease state related to infection with *Borrelia* has already been made, the present invention is useful for monitoring progression or regression of the disease state whereby patients exhibiting enhanced *Borrelia* gene expression will experience a worse clinical outcome relative to patients expressing these gene(s) at a lower level.

By "biological sample" is intended any biological sample obtained from an animal, cell line, tissue culture, or other source which contains *Borrelia* polypeptide, mRNA, or DNA. Biological samples include body fluids (such as saliva, blood, plasma, urine, mucus, synovial fluid, etc.) tissues (such as muscle, skin, and cartilage) and any other biological source suspected of containing *Borrelia* polypeptides or nucleic acids. Methods for obtaining biological samples such as tissue are well known in the art.

The present invention is useful for detecting diseases related to *Borrelia* infections in animals. Preferred animals include monkeys, apes, cats, dogs, birds, cows, pigs, mice, horses, rabbits and humans. Particularly preferred are humans.

Total RNA can be isolated from a biological sample using any suitable technique such as the single-step guanidinium-thiocyanate-phenol-chloroform method described in Chomczynski et al. (1987) Anal. Biochem. 162:156-159. mRNA encoding *Borrelia* polypeptides having sufficient homology to the nucleic acid sequences identified in SEQ ID NOS:1-744 to allow for hybridization between complementary sequences are then assayed using any appropriate method. These include Northern blot analysis, S1 nuclease mapping, the polymerase chain reaction (PCR), reverse transcription in combination with the polymerase chain reaction (RT-PCR), and reverse transcription in combination with the ligase chain reaction (RT-LCR).

Northern blot analysis can be performed as described in Harada et al. (1990) Cell 63:303-312. Briefly, total RNA is prepared from a biological sample as described above. For the Northern blot, the RNA is denatured in an appropriate buffer (such as glyoxal/dimethyl sulfoxide/sodium phosphate buffer), subjected to agarose gel electrophoresis, and transferred onto a nitrocellulose filter. After the RNAs have been linked to the filter by a UV linker, the filter is prehybridized in a solution containing formamide, SSC, Denhardt's solution, denatured

salmon sperm, SDS, and sodium phosphate buffer. A *T. pallidum* polynucleotide sequence shown in SEQ ID NOS:1-744, or portion thereof, labeled according to any appropriate method (such as the ^{32}P-multiprimed DNA labeling system (Amersham)) is used as probe. After hybridization overnight, the filter is washed and exposed to x-ray film. DNA for use as probe according to the present invention is described in the sections above and will preferably at least 15 nucleotides in length.

S1 mapping can be performed as described in Fujita et al. (1987) Cell 49:357-367. To prepare probe DNA for use in S1 mapping, the sense strand of an above-described *T. pallidum* DNA sequence of the present invention is used as a template to synthesize labeled antisense DNA. The antisense DNA can then be digested using an appropriate restriction endonuclease to generate further DNA probes of a desired length. Such antisense probes are useful for visualizing protected bands corresponding to the target mRNA (*i.e.*, mRNA encoding *Borrelia* polypeptides).

Levels of mRNA encoding *Borrelia* polypeptides are assayed, for *e.g.*, using the RT-PCR method described in Makino et al. (1990) Technique 2:295-301. By this method, the radioactivities of the "amplicons" in the polyacrylamide gel bands are linearly related to the initial concentration of the target mRNA. Briefly, this method involves adding total RNA isolated from a biological sample in a reaction mixture containing a RT primer and appropriate buffer. After incubating for primer annealing, the mixture can be supplemented with a RT buffer, dNTPs, DTT, RNase inhibitor and reverse transcriptase. After incubation to achieve reverse transcription of the RNA, the RT products are then subject to PCR using labeled primers. Alternatively, rather than labeling the primers, a labeled dNTP can be included in the PCR reaction mixture. PCR amplification can be performed in a DNA thermal cycler according to conventional techniques. After a suitable number of rounds to achieve amplification, the PCR reaction mixture is electrophoresed on a polyacrylamide gel. After drying the gel, the radioactivity of the appropriate bands (corresponding to the mRNA encoding the *Borrelia* polypeptides of the present invention) are quantified using an imaging analyzer. RT and PCR reaction ingredients and conditions, reagent and gel concentrations, and labeling methods are well known in the art. Variations on the RT-PCR method will be apparent to the skilled artisan. Other PCR methods that can detect the nucleic acid of the present invention can be found in PCR PRIMER: A LABORATORY MANUAL (C.W. Dieffenbach et al. eds., Cold Spring Harbor Lab Press, 1995).

The polynucleotides of the present invention, including both DNA and RNA, may be used to detect polynucleotides of the present invention or Borrelia species including *T. pallidum* using bio chip technology. The present invention includes both high density chip arrays (>1000 oligonucleotides per cm^2) and low density chip arrays (<1000 oligonucleotides per cm^2). Bio chips comprising arrays of polynucleotides of the present invention may be used to detect Borrelia species, including *T. pallidum*, in biological and environmental samples and to diagnose an animal, including humans, with an *T. pallidum* or other Borrelia infection. The bio chips of the present invention may comprise polynucleotide sequences of other pathogens including

bacteria, viral, parasitic, and fungal polynucleotide sequences, in addition to the polynucleotide sequences of the present invention, for use in rapid diffenertial pathogenic detection and diagnosis. The bio chips can also be used to monitor an *T. pallidum* or other Borrelia infections and to monitor the genetic changes (deletions, insertions, mismatches, etc.) in response to drug therapy in the clinic and drug development in the laboratory. The bio chip technology comprising arrays of polynucleotides of the present invention may also be used to simultaneously monitor the expression of a multiplicity of genes, including those of the present invention. The polynucleotides used to comprise a selected array may be specified in the same manner as for the fragements, i.e, by their 5' and 3' positions or length in contigious base pairs and include from. Methods and particular uses of the polynucleotides of the present invention to detect Borrelia species, including *T. pallidum*, using bio chip technology include those known in the art and those of: U.S. Patent Nos. 5510270, 5545531, 5445934, 5677195, 5532128, 5556752, 5527681, 5451683, 5424186, 5607646, 5658732 and World Patent Nos. WO/9710365, WO/9511995, WO/9743447, WO/9535505, each incorporated herein in their entireties.

Biosensors using the polynucleotides of the present invention may also be used to detect, diagnose, and monitor *T. pallidum* or other Borrelia species and infections thereof. Biosensors using the polynucleotides of the present invention may also be used to detect particular polynucleotides of the present invention. Biosensors using the polynucleotides of the present invention may also be used to monitor the genetic changes (deletions, insertions, mismatches, etc.) in response to drug therapy in the clinic and drug development in the laboratory. Methods and particular uses of the polynucleotides of the present invention to detect Borrelia species, including *T. pallidum*, using biosenors include those known in the art and those of: U.S. Patent Nos 5721102, 5658732, 5631170, and World Patent Nos. WO97/35011, WO/9720203, each incorporated herein in their entireties.

Thus, the present invention includes both bio chips and biosensors comprising polynucleotides of the present invention and methods of their use.

Assaying *Borrelia* polypeptide levels in a biological sample can occur using any art-known method, such as antibody-based techniques. For example, *Borrelia* polypeptide expression in tissues can be studied with classical immunohistological methods. In these, the specific recognition is provided by the primary antibody (polyclonal or monoclonal) but the secondary detection system can utilize fluorescent, enzyme, or other conjugated secondary antibodies. As a result, an immunohistological staining of tissue section for pathological examination is obtained. Tissues can also be extracted, *e.g.*, with urea and neutral detergent, for the liberation of *Borrelia* polypeptides for Western-blot or dot/slot assay. *See, e.g.*, Jalkanen, M. et al. (1985) J. Cell. Biol. 101:976-985; Jalkanen, M. et al. (1987) J. Cell . Biol. 105:3087-3096. In this technique, which is based on the use of cationic solid phases, quantitation of a *Borrelia* polypeptide can be accomplished using an isolated *Borrelia* polypeptide as a standard. This technique can also be applied to body fluids.

Other antibody-based methods useful for detecting *Borrelia* polypeptide gene expression include immunoassays, such as the ELISA and the radioimmunoassay (RIA). For example, a *Borrelia* polypeptide-specific monoclonal antibodies can be used both as an immunoabsorbent and as an enzyme-labeled probe to detect and quantify a *Borrelia* polypeptide. The amount of a *Borrelia* polypeptide present in the sample can be calculated by reference to the amount present in a standard preparation using a linear regression computer algorithm. Such an ELISA is described in Iacobelli et al. (1988) Breast Cancer Research and Treatment 11:19-30. In another ELISA assay, two distinct specific monoclonal antibodies can be used to detect *Borrelia* polypeptides in a body fluid. In this assay, one of the antibodies is used as the immunoabsorbent and the other as the enzyme-labeled probe.

The above techniques may be conducted essentially as a "one-step" or "two-step" assay. The "one-step" assay involves contacting the *Borrelia* polypeptide with immobilized antibody and, without washing, contacting the mixture with the labeled antibody. The "two-step" assay involves washing before contacting the mixture with the labeled antibody. Other conventional methods may also be employed as suitable. It is usually desirable to immobilize one component of the assay system on a support, thereby allowing other components of the system to be brought into contact with the component and readily removed from the sample. Variations of the above and other immunological methods included in the present invention can also be found in Harlow et al., ANTIBODIES: A LABORATORY MANUAL, (Cold Spring Harbor Laboratory Press, 2nd ed. 1988).

Suitable enzyme labels include, for example, those from the oxidase group, which catalyze the production of hydrogen peroxide by reacting with substrate. Glucose oxidase is particularly preferred as it has good stability and its substrate (glucose) is readily available. Activity of an oxidase label may be assayed by measuring the concentration of hydrogen peroxide formed by the enzyme-labeled antibody/substrate reaction. Besides enzymes, other suitable labels include radioisotopes, such as iodine (^{125}I, ^{121}I), carbon (^{14}C), sulphur (^{35}S), tritium (^{3}H), indium (^{112}In), and technetium (^{99m}Tc), and fluorescent labels, such as fluorescein and rhodamine, and biotin.

Further suitable labels for the *Borrelia* polypeptide-specific antibodies of the present invention are provided below. Examples of suitable enzyme labels include malate dehydrogenase, Borrelia nuclease, delta-5-steroid isomerase, yeast-alcohol dehydrogenase, alpha-glycerol phosphate dehydrogenase, triose phosphate isomerase, peroxidase, alkaline phosphatase, asparaginase, glucose oxidase, beta-galactosidase, ribonuclease, urease, catalase, glucose-6-phosphate dehydrogenase, glucoamylase, and acetylcholine esterase.

Examples of suitable radioisotopic labels include ^{3}H, ^{111}In, ^{125}I, ^{131}I, ^{32}P, ^{35}S, ^{14}C, ^{51}Cr, ^{57}To, ^{58}Co, ^{59}Fe, ^{75}Se, ^{152}Eu, ^{90}Y, ^{67}Cu, ^{217}Ci, ^{211}At, ^{212}Pb, ^{47}Sc, ^{109}Pd, etc. ^{111}In is a preferred isotope where *in vivo* imaging is used since its avoids the problem of dehalogenation of the ^{125}I or ^{131}I-labeled monoclonal antibody by the liver. In addition, this radionucleotide has a more favorable gamma emission energy for imaging. *See, e.g.,* Perkins et al. (1985) Eur. J. Nucl.

Med. 10:296-301; Carasquillo et al. (1987) J. Nucl. Med. 28:281-287. For example, ^{111}In coupled to monoclonal antibodies with 1-(P-isothiocyanatobenzyl)-DPTA has shown little uptake in non-tumors tissues, particularly the liver, and therefore enhances specificity of tumor localization. See, Esteban et al. (1987) J. Nucl. Med. 28:861-870.

Examples of suitable non-radioactive isotopic labels include ^{157}Gd, ^{55}Mn, ^{162}Dy, ^{52}Tr, and ^{56}Fe.

Examples of suitable fluorescent labels include an ^{152}Eu label, a fluorescein label, an isothiocyanate label, a rhodamine label, a phycoerythrin label, a phycocyanin label, an allophycocyanin label, an o-phthaldehyde label, and a fluorescamine label.

Examples of suitable toxin labels include, *Pseudomonas* toxin, diphtheria toxin, ricin, and cholera toxin.

Examples of chemiluminescent labels include a luminal label, an isoluminal label, an aromatic acridinium ester label, an imidazole label, an acridinium salt label, an oxalate ester label, a luciferin label, a luciferase label, and an aequorin label.

Examples of nuclear magnetic resonance contrasting agents include heavy metal nuclei such as Gd, Mn, and iron.

Typical techniques for binding the above-described labels to antibodies are provided by Kennedy et al. (1976) Clin. Chim. Acta 70:1-31, and Schurs et al. (1977) Clin. Chim. Acta 81:1-40. Coupling techniques mentioned in the latter are the glutaraldehyde method, the periodate method, the dimaleimide method, the m-maleimidobenzyl-N-hydroxy-succinimide ester method, all of which methods are incorporated by reference herein.

In a related aspect, the invention includes a diagnostic kit for use in screening serum containing antibodies specific against *T. pallidum* infection. Such a kit may include an isolated *T. pallidum* antigen comprising an epitope which is specifically immunoreactive with at least one anti-*T. pallidum* antibody. Such a kit also includes means for detecting the binding of said antibody to the antigen. In specific embodiments, the kit may include a recombinantly produced or chemically synthesized peptide or polypeptide antigen. The peptide or polypeptide antigen may be attached to a solid support.

In a more specific embodiment, the detecting means of the above-described kit includes a solid support to which said peptide or polypeptide antigen is attached. Such a kit may also include a non-attached reporter-labeled anti-human antibody. In this embodiment, binding of the antibody to the *T. pallidum* antigen can be detected by binding of the reporter labeled antibody to the anti-*T. pallidum* polypeptide antibody.

Specifically, the invention provides a compartmentalized kit to receive, in close confinement, one or more containers which comprises: (a) a first container comprising one of the DFs or antibodies of the present invention; and (b) one or more other containers comprising one or more of the following: wash reagents, reagents capable of detecting presence of a bound DF or antibody.

In detail, a compartmentalized kit includes any kit in which reagents are contained in separate containers. Such containers include small glass containers, plastic containers or strips of plastic or paper. Such containers allows one to efficiently transfer reagents from one compartment to another compartment such that the samples and reagents are not cross-contaminated, and the agents or solutions of each container can be added in a quantitative fashion from one compartment to another. Such containers will include a container which will accept the test sample, a container which contains the antibodies used in the assay, containers which contain wash reagents (such as phosphate buffered saline, Tris-buffers, *etc.*), and containers which contain the reagents used to detect the bound antibody or DF.

In a related aspect, the invention includes a method of detecting *T. pallidum* infection in a subject. This detection method includes reacting a body fluid, preferably serum, from the subject with an isolated *T. pallidum* antigen, and examining the antigen for the presence of bound antibody. In a specific embodiment, the method includes a polypeptide antigen attached to a solid support, and serum is reacted with the support. Subsequently, the support is reacted with a reporter-labeled anti-human antibody. The support is then examined for the presence of reporter-labeled antibody.

The solid surface reagent employed in the above assays and kits is prepared by known techniques for attaching protein material to solid support material, such as polymeric beads, dip sticks, 96-well plates or filter material. These attachment methods generally include non-specific adsorption of the protein to the support or covalent attachment of the protein , typically through a free amine group, to a chemically reactive group on the solid support, such as an activated carboxyl, hydroxyl, or aldehyde group. Alternatively, streptavidin coated plates can be used in conjunction with biotinylated antigen(s).

The polypeptides and antibodies of the present invention, including fragments thereof, may be used to detect Borrelia species including *T. pallidum* using bio chip and biosensor technology. Bio chip and biosensors of the present invention may comprise the polypeptides of the present invention to detect antibodies, which specifically recognize Borrelia species, including *T. pallidum*. Bio chip and biosensors of the present invention may also comprise antibodies which specifically recognize the polypeptides of the present invention to detect Borrelia species, including *T. pallidum* or specific polypeptides of the present invention. Bio chips or biosensors comprising polypeptides or antibodies of the present invention may be used to detect Borrelia species, including *T. pallidum*, in biological and environmental samples and to diagnose an animal, including humans, with an *T. pallidum* or other Borrelia infection. Thus, the present invention includes both bio chips and biosensors comprising polypeptides or antibodies of the present invention and methods of their use.

The bio chips of the present invention may further comprise polypeptide sequences of other pathogens including bacteria, viral, parasitic, and fungal polypeptide sequences, in addition to the polypeptide sequences of the present invention, for use in rapid diffenertial pathogenic

detection and diagnosis. The bio chips of the present invention may further comprise antibodies or fragements thereof specific for other pathogens including bacteria, viral, parasitic, and fungal polypeptide sequences, in addition to the antibodies or fragements thereof of the present invention, for use in rapid diffenertial pathogenic detection and diagnosis. The bio chips and biosensors of the present invention may also be used to monitor an *T. pallidum* or other Borrelia infection and to monitor the genetic changes (amio acid deletions, insertions, substitutions, etc.) in response to drug therapy in the clinic and drug development in the laboratory. The bio chip and biosensors comprising polypeptides or antibodies of the present invention may also be used to simultaneously monitor the expression of a multiplicity of polypeptides, including those of the present invention. The polypeptides used to comprise a bio chip or biosensor of the present invention may be specified in the same manner as for the fragements, i.e, by their N-terminal and C-terminal positions or length in contigious amino acid residue. Methods and particular uses of the polypeptides and antibodies of the present invention to detect Borrelia species, including *T. pallidum*, or specific polypeptides using bio chip and biosensor technology include those known in the art, those of the U.S. Patent Nos. and World Patent Nos. listed above for bio chips and biosensors using polynucleotides of the present invention, and those of: U.S. Patent Nos. 5658732, 5135852, 5567301, 5677196, 5690894 and World Patent Nos. WO9729366, WO9612957, each incorporated herein in their entireties.

4. Screening Assay for Binding Agents

Using the isolated proteins of the present invention, the present invention further provides methods of obtaining and identifying agents which bind to a protein encoded by one of the ORFs of the present invention or to one of the fragments and the *T. pallidum* fragment and contigs herein described.

In general, such methods comprise steps of:

(a) contacting an agent with an isolated protein encoded by one of the ORFs of the present invention, or an isolated fragment of the *T. pallidum* genome; and

(b) determining whether the agent binds to said protein or said fragment.

The agents screened in the above assay can be, but are not limited to, peptides, carbohydrates, vitamin derivatives, or other pharmaceutical agents. The agents can be selected and screened at random or rationally selected or designed using protein modeling techniques.

For random screening, agents such as peptides, carbohydrates, pharmaceutical agents and the like are selected at random and are assayed for their ability to bind to the protein encoded by the ORF of the present invention.

Alternatively, agents may be rationally selected or designed. As used herein, an agent is said to be "rationally selected or designed" when the agent is chosen based on the configuration of the particular protein. For example, one skilled in the art can readily adapt currently available procedures to generate peptides, pharmaceutical agents and the like capable of binding to a specific peptide sequence in order to generate rationally designed antipeptide peptides, for

example see Hurby *et al.*, "Application of Synthetic Peptides: Antisense Peptides," in *Synthetic Peptides, A User's Guide*, W. H. Freeman, NY (1992), pp. 289-307, and Kaspczak *et al.*, *Biochemistry 28:*9230-8 (1989), or pharmaceutical agents, or the like.

In addition to the foregoing, one class of agents of the present invention, as broadly described, can be used to control gene expression through binding to one of the ORFs or EMFs of the present invention. As described above, such agents can be randomly screened or rationally designed/selected. Targeting the ORF or EMF allows a skilled artisan to design sequence specific or element specific agents, modulating the expression of either a single ORF or multiple ORFs which rely on the same EMF for expression control.

One class of DNA binding agents are agents which contain base residues which hybridize or form a triple helix by binding to DNA or RNA. Such agents can be based on the classic phosphodiester, ribonucleic acid backbone, or can be a variety of sulfhydryl or polymeric derivatives which have base attachment capacity.

Agents suitable for use in these methods usually contain 20 to 40 bases and are designed to be complementary to a region of the gene involved in transcription (triple helix - see Lee *et al.*, *Nucl. Acids Res. 6:*3073 (1979); Cooney *et al.*, *Science 241:*456 (1988); and Dervan *et al.*, *Science 251:*1360 (1991)) or to the mRNA itself (antisense - Okano, *J. Neurochem. 56:*560 (1991); *Oligodeoxynucleotides as Antisense Inhibitors of Gene Expression,* CRC Press, Boca Raton, FL (1988)). Triple helix- formation optimally results in a shut-off of RNA transcription from DNA, while antisense RNA hybridization blocks translation of an mRNA molecule into polypeptide. Both techniques have been demonstrated to be effective in model systems. Information contained in the sequences of the present invention can be used to design antisense and triple helix-forming oligonucleotides, and other DNA binding agents.

5. Pharmaceutical Compositions and Vaccines

The present invention further provides pharmaceutical agents which can be used to modulate the growth or pathogenicity of *T. pallidum*, or another related organism, *in vivo* or *in vitro*. As used herein, a "pharmaceutical agent" is defined as a composition of matter which can be formulated using known techniques to provide a pharmaceutical compositions. As used herein, the "pharmaceutical agents of the present invention" refers the pharmaceutical agents which are derived from the proteins encoded by the ORFs of the present invention or are agents which are identified using the herein described assays.

As used herein, a pharmaceutical agent is said to "modulate the growth pathogenicity of *T. pallidum* or a related organism, *in vivo* or *in vitro*," when the agent reduces the rate of growth, rate of division, or viability of the organism in question. The pharmaceutical agents of the present invention can modulate the growth or pathogenicity of an organism in many fashions, although an understanding of the underlying mechanism of action is not needed to practice the use of the pharmaceutical agents of the present invention. Some agents will modulate the growth by binding to an important protein thus blocking the biological activity of the protein, while other

agents may bind to a component of the outer surface of the organism blocking attachment or rendering the organism more prone to act the bodies nature immune system. Alternatively, the agent may comprise a protein encoded by one of the ORFs of the present invention and serve as a vaccine. The development and use of a vaccine based on outer membrane components are well known in the art.

As used herein, a "related organism" is a broad term which refers to any organism whose growth can be modulated by one of the pharmaceutical agents of the present invention. In general, such an organism will contain a homolog of the protein which is the target of the pharmaceutical agent or the protein used as a vaccine. As such, related organisms do not need to be bacterial but may be fungal or viral pathogens.

The pharmaceutical agents and compositions of the present invention may be administered in a convenient manner, such as by the oral, topical, intravenous, intraperitoneal, intramuscular, subcutaneous, intranasal or intradermal routes. The pharmaceutical compositions are administered in an amount which is effective for treating and/or prophylaxis of the specific indication. In general, they are administered in an amount of at least about 1 mg/kg body weight and in most cases they will be administered in an amount not in excess of about 1 g/kg body weight per day. In most cases, the dosage is from about 0.1 mg/kg to about 10 g/kg body weight daily, taking into account the routes of administration, symptoms, *etc.*

The agents of the present invention can be used in native form or can be modified to form a chemical derivative. As used herein, a molecule is said to be a "chemical derivative" of another molecule when it contains additional chemical moieties not normally a part of the molecule. Such moieties may improve the molecule's solubility, absorption, biological half life, *etc.* The moieties may alternatively decrease the toxicity of the molecule, eliminate or attenuate any undesirable side effect of the molecule, *etc.* Moieties capable of mediating such effects are disclosed in, among other sources, REMINGTON'S PHARMACEUTICAL SCIENCES (1980) cited elsewhere herein.

For example, such moieties may change an immunological character of the functional derivative, such as affinity for a given antibody. Such changes in immunomodulation activity are measured by the appropriate assay, such as a competitive type immunoassay. Modifications of such protein properties as redox or thermal stability, biological half-life, hydrophobicity, susceptibility to proteolytic degradation or the tendency to aggregate with carriers or into multimers also may be effected in this way and can be assayed by methods well known to the skilled artisan.

The therapeutic effects of the agents of the present invention may be obtained by providing the agent to a patient by any suitable means (*e.g.*, inhalation, intravenously, intramuscularly, subcutaneously, enterally, or parenterally). It is preferred to administer the agent of the present invention so as to achieve an effective concentration within the blood or tissue in which the growth of the organism is to be controlled. To achieve an effective blood

concentration, the preferred method is to administer the agent by injection. The administration may be by continuous infusion, or by single or multiple injections.

In providing a patient with one of the agents of the present invention, the dosage of the administered agent will vary depending upon such factors as the patient's age, weight, height, sex, general medical condition, previous medical history, *etc.* In general, it is desirable to provide the recipient with a dosage of agent which is in the range of from about 1 pg/kg to 10 mg/kg (body weight of patient), although a lower or higher dosage may be administered. The therapeutically effective dose can be lowered by using combinations of the agents of the present invention or another agent.

As used herein, two or more compounds or agents are said to be administered "in combination" with each other when either (1) the physiological effects of each compound, or (2) the serum concentrations of each compound can be measured at the same time. The composition of the present invention can be administered concurrently with, prior to, or following the administration of the other agent.

The agents of the present invention are intended to be provided to recipient subjects in an amount sufficient to decrease the rate of growth (as defined above) of the target organism.

The administration of the agent(s) of the invention may be for either a "prophylactic" or "therapeutic" purpose. When provided prophylactically, the agent(s) are provided in advance of any symptoms indicative of the organisms growth. The prophylactic administration of the agent(s) serves to prevent, attenuate, or decrease the rate of onset of any subsequent infection. When provided therapeutically, the agent(s) are provided at (or shortly after) the onset of an indication of infection. The therapeutic administration of the compound(s) serves to attenuate the pathological symptoms of the infection and to increase the rate of recovery.

The agents of the present invention are administered to a subject, such as a mammal, or a patient, in a pharmaceutically acceptable form and in a therapeutically effective concentration. A composition is said to be "pharmacologically acceptable" if its administration can be tolerated by a recipient patient. Such an agent is said to be administered in a "therapeutically effective amount" if the amount administered is physiologically significant. An agent is physiologically significant if its presence results in a detectable change in the physiology of a recipient patient.

The agents of the present invention can be formulated according to known methods to prepare pharmaceutically useful compositions, whereby these materials, or their functional derivatives, are combined in a mixture with a pharmaceutically acceptable carrier vehicle. Suitable vehicles and their formulation, inclusive of other human proteins, *e.g.*, human serum albumin, are described, for example, in REMINGTON'S PHARMACEUTICAL SCIENCES, 16th Ed., Osol, A., Ed., Mack Publishing, Easton PA (1980). In order to form a pharmaceutically acceptable composition suitable for effective administration, such compositions will contain an effective amount of one or more of the agents of the present invention, together with a suitable amount of carrier vehicle.

Additional pharmaceutical methods may be employed to control the duration of action. Control release preparations may be achieved through the use of polymers to complex or absorb one or more of the agents of the present invention. The controlled delivery may be effectuated by a variety of well known techniques, including formulation with macromolecules such as, for example, polyesters, polyamino acids, polyvinyl, pyrrolidone, ethylenevinylacetate, methylcellulose, carboxymethylcellulose, or protamine, sulfate, adjusting the concentration of the macromolecules and the agent in the formulation, and by appropriate use of methods of incorporation, which can be manipulated to effectuate a desired time course of release. Another possible method to control the duration of action by controlled release preparations is to incorporate agents of the present invention into particles of a polymeric material such as polyesters, polyamino acids, hydrogels, poly(lactic acid) or ethylene vinylacetate copolymers. Alternatively, instead of incorporating these agents into polymeric particles, it is possible to entrap these materials in microcapsules prepared, for example, by coacervation techniques or by interfacial polymerization with, for example, hydroxymethylcellulose or gelatine-microcapsules and poly(methylmethacylate) microcapsules, respectively, or in colloidal drug delivery systems, for example, liposomes, albumin microspheres, microemulsions, nanoparticles, and nanocapsules or in macroemulsions. Such techniques are disclosed in REMINGTON'S PHARMACEUTICAL SCIENCES (1980).

The invention further provides a pharmaceutical pack or kit comprising one or more containers filled with one or more of the ingredients of the pharmaceutical compositions of the invention. Associated with such container(s) can be a notice in the form prescribed by a governmental agency regulating the manufacture, use or sale of pharmaceuticals or biological products, which notice reflects approval by the agency of manufacture, use or sale for human administration.

In addition, the agents of the present invention may be employed in conjunction with other therapeutic compounds.

6. Shot-Gun Approach to Megabase DNA Sequencing

The present invention further demonstrates that a large sequence can be sequenced using a random shotgun approach. This procedure, described in detail in the examples that follow, has eliminated the up front cost of isolating and ordering overlapping or contiguous subclones prior to the start of the sequencing protocols.

Certain aspects of the present invention are described in greater detail in the examples that follow. The examples are provided by way of illustration. Other aspects and embodiments of the present invention are contemplated by the inventors, as will be clear to those of skill in the art from reading the present disclosure.

ILLUSTRATIVE EXAMPLES

LIBRARIES AND SEQUENCING

1. Shotgun Sequencing Probability Analysis

The overall strategy for a shotgun approach to whole genome sequencing follows from the Lander and Waterman (Landerman and Waterman, *Genomics* 2:231 (1988)) application of the equation for the Poisson distribution. According to this treatment, the probability, P0, that any given base in a sequence of size L, in nucleotides, is not sequenced after a certain amount, n, in nucleotides, of random sequence has been determined can be calculated by the equation P0 = e-m, where m is L/n, the fold coverage. For instance, for a genome of 2.8 Mb, m=1 when 2.8 Mb of sequence has been randomly generated (1X coverage). At that point, P0 = e-1 = 0.37. The probability that any given base has not been sequenced is the same as the probability that any region of the whole sequence L has not been determined and, therefore, is equivalent to the fraction of the whole sequence that has yet to be determined. Thus, at one-fold coverage, approximately 37% of a polynucleotide of size L, in nucleotides has not been sequenced. When 14 Mb of sequence has been generated, coverage is 5X for a 2.8 Mb and the unsequenced fraction drops to .0067 or 0.67%. 5X coverage of a 2.8 Mb sequence can be attained by sequencing approximately 17,000 random clones from both insert ends with an average sequence read length of 410 bp.

Similarly, the total gap length, G, is determined by the equation G = Le-m, and the average gap size, g, follows the equation, g = L/n. Thus, 5X coverage leaves about 240 gaps averaging about 82 bp in size in a sequence of a polynucleotide 2.8 Mb long.

The treatment above is essentially that of Lander and Waterman, *Genomics* 2: 231 (1988).

2. Random Library Construction

In order to approximate the random model described above during actual sequencing, a nearly ideal library of cloned genomic fragments is required. The following library construction procedure was developed to achieve this end.

T. pallidum DNA is prepared by phenol extraction. A mixture containing 200 µg DNA in 1.0 ml of 300 mM sodium acetate, 10 mM Tris-HCl, 1 mM Na-EDTA, 50% glycerol is processed through a nebulizer (IPI Medical Products) with a stream of nitrogen adjusted to 35 Kpa for 2 minutes. The sonicated DNA is ethanol precipitated and redissolved in 500 µl TE buffer.

To create blunt-ends, a 100 µl aliquot of the resuspended DNA is digested with 5 units of BAL31 nuclease (New England BioLabs) for 10 min at 30°C in 200 µl BAL31 buffer. The digested DNA is phenol-extracted, ethanol-precipitated, redissolved in 100 µl TE buffer, and then size-fractionated by electrophoresis through a 1.0% low melting temperature agarose gel. The section containing DNA fragments 1.6-2.0 kb in size is excised from the gel, and the LGT agarose is melted and the resulting solution is extracted with phenol to separate the agarose from

the DNA. DNA is ethanol precipitated and redissolved in 20 µl of TE buffer for ligation to vector.

A two-step ligation procedure is used to produce a plasmid library with 97% inserts, of which >99% were single inserts. The first ligation mixture (50 ul) contains 2 µg of DNA fragments, 2 µg pUC18 DNA (Pharmacia) cut with SmaI and dephosphorylated with bacterial alkaline phosphatase, and 10 units of T4 ligase (GIBCO/BRL) and is incubated at 14°C for 4 hr. The ligation mixture then is phenol extracted and ethanol precipitated, and the precipitated DNA is dissolved in 20 µl TE buffer and electrophoresed on a 1.0% low melting agarose gel. Discrete bands in a ladder are visualized by ethidium bromide-staining and UV illumination and identified by size as insert (I), vector (v), v+I, v+2i, v+3i, *etc.* The portion of the gel containing v+I DNA is excised and the v+I DNA is recovered and resuspended into 20 µl TE. The v+I DNA then is blunt-ended by T4 polymerase treatment for 5 min. at 37°C in a reaction mixture (50 ul) containing the v+I linears, 500 µM each of the 4 dNTPs, and 9 units of T4 polymerase (New England BioLabs), under recommended buffer conditions. After phenol extraction and ethanol precipitation the repaired v+I linears are dissolved in 20 µl TE. The final ligation to produce circles is carried out in a 50 µl reaction containing 5 µl of v+I linears and 5 units of T4 ligase at 14°C overnight. After 10 min. at 70°C the following day, the reaction mixture is stored at -20°C.

This two-stage procedure results in a molecularly random collection of single-insert plasmid recombinants with minimal contamination from double-insert chimeras (<1%) or free vector (<3%).

Since deviation from randomness can arise from propagation the DNA in the host, *E. coli* host cells deficient in all recombination and restriction functions (A. Greener, *Strategies 3 (1)*:5 (1990)) are used to prevent rearrangements, deletions, and loss of clones by restriction. Furthermore, transformed cells are plated directly on antibiotic diffusion plates to avoid the usual broth recovery phase which allows multiplication and selection of the most rapidly growing cells.

Plating is carried out as follows. A 100 µl aliquot of Epicurian Coli SURE II Supercompetent Cells (Stratagene 200152) is thawed on ice and transferred to a chilled Falcon 2059 tube on ice. A 1.7 µl aliquot of 1.42 M beta-mercaptoethanol is added to the aliquot of cells to a final concentration of 25 mM. Cells are incubated on ice for 10 min. A 1 µl aliquot of the final ligation is added to the cells and incubated on ice for 30 min. The cells are heat pulsed for 30 sec. at 42°C and placed back on ice for 2 min. The outgrowth period in liquid culture is eliminated from this protocol in order to minimize the preferential growth of any given transformed cell. Instead the transformation mixture is plated directly on a nutrient rich SOB plate containing a 5 ml bottom layer of SOB agar (5% SOB agar: 20 g tryptone, 5 g yeast extract, 0.5 g NaCl, 1.5% Difco Agar per liter of media). The 5 ml bottom layer is supplemented with 0.4 ml of 50 mg/ml ampicillin per 100 ml SOB agar. The 15 ml top layer of SOB agar is supplemented with 1 ml X-Gal (2%), 1 ml MgCl2 (1 M), and 1 ml MgSO4/100 ml SOB agar. The 15 ml top layer is poured just prior to plating. Our titer is approximately 100 colonies/10 µl aliquot of transformation.

All colonies are picked for template preparation regardless of size. Thus, only clones lost due to "poison" DNA or deleterious gene products are deleted from the library, resulting in a slight increase in gap number over that expected.

3. Random DNA Sequencing

High quality double stranded DNA plasmid templates are prepared using a "boiling bead" method developed in collaboration with Advanced Genetic Technology Corp. (Gaithersburg, MD) (Adams *et al.*, *Science 252*:1651 (1991); Adams *et al.*, *Nature 355*:632 (1992)). Plasmid preparation is performed in a 96-well format for all stages of DNA preparation from bacterial growth through final DNA purification. Template concentration is determined using Hoechst Dye and a Millipore Cytofluor. DNA concentrations are not adjusted, but low-yielding templates are identified where possible and not sequenced.

Templates are also prepared from two *T. pallidum* lambda genomic libraries. An amplified library is constructed in the vector Lambda GEM-12 (Promega) and an unamplified library is constructed in Lambda DASH II (Stratagene). In particular, for the unamplified lambda library, *T. pallidum* DNA (> 100 kb) is partially digested in a reaction mixture (200 ul) containing 50 µg DNA, 1X Sau3AI buffer, 20 units Sau3AI for 6 min. at 23°C. The digested DNA was phenol-extracted and electrophoresed on a 0.5% low melting agarose gel at 2V/cm for 7 hours. Fragments from 15 to 25 kb are excised and recovered in a final volume of 6 ul. One µl of fragments is used with 1 µl of DASHII vector (Stratagene) in the recommended ligation reaction. One µl of the ligation mixture is used per packaging reaction following the recommended protocol with the Gigapack II XL Packaging Extract (Stratagene, #227711). Phage are plated directly without amplification from the packaging mixture (after dilution with 500 µl of recommended SM buffer and chloroform treatment). Yield is about 2.5x103 pfu/ul. The amplified library is prepared essentially as above except the lambda GEM-12 vector is used. After packaging, about 3.5x104 pfu are plated on the restrictive NM539 host. The lysate is harvested in 2 ml of SM buffer and stored frozen in 7% dimethylsulfoxide. The phage titer is approximately 1x109 pfu/ml.

Liquid lysates (100 µl) are prepared from randomly selected plaques (from the unamplified library) and template is prepared by long-range PCR using T7 and T3 vector-specific primers.

Sequencing reactions are carried out on plasmid and/or PCR templates using the AB Catalyst LabStation with Applied Biosystems PRISM Ready Reaction Dye Primer Cycle Sequencing Kits for the M13 forward (M13-21) and the M13 reverse (M13RP1) primers (Adams *et al.*, *Nature 368*:474 (1994)). Dye terminator sequencing reactions are carried out on the lambda templates on a Perkin-Elmer 9600 Thermocycler using the Applied Biosystems Ready Reaction Dye Terminator Cycle Sequencing kits. T7 and SP6 primers are used to sequence the ends of the inserts from the Lambda GEM-12 library and T7 and T3 primers are used to sequence the ends of the inserts from the Lambda DASH II library. Sequencing reactions are performed

by eight individuals using an average of fourteen AB 373 DNA Sequencers per day. All sequencing reactions are analyzed using the Stretch modification of the AB 373, primarily using a 34 cm well-to-read distance. The overall sequencing success rate very approximately is about 85% for M13-21 and M13RP1 sequences and 65% for dye-terminator reactions. The average usable read length is 485 bp for M13-21 sequences, 445bp for M13RP1 sequences, and 375 bp for dye-terminator reactions.

Richards *et al.*, Chapter 28 in AUTOMATED DNA SEQUENCING AND ANALYSIS, M. D. Adams, C. Fields, J. C. Venter, Eds., Academic Press, London, (1994) described the value of using sequence from both ends of sequencing templates to facilitate ordering of contigs in shotgun assembly projects of lambda and cosmid clones. We balance the desirability of both-end sequencing (including the reduced cost of lower total number of templates) against shorter read-lengths for sequencing reactions performed with the M13RP1 (reverse) primer compared to the M13-21 (forward) primer. Approximately one-half of the templates are sequenced from both ends. Random reverse sequencing reactions are done based on successful forward sequencing reactions. Some M13RP1 sequences are obtained in a semi-directed fashion: M13-21: sequences pointing outward at the ends of contigs are chosen for M13RP1 sequencing in an effort to specifically order contigs.

4. Protocol for Automated Cycle Sequencing

The sequencing is carried out using ABI Catalyst robots and AB 373 Automated DNA Sequencers. The Catalyst robot is a publicly available sophisticated pipetting and temperature control robot which has been developed specifically for DNA sequencing reactions. The Catalyst combines pre-aliquoted templates and reaction mixes consisting of deoxy- and dideoxynucleotides, the thermostable Taq DNA polymerase, fluorescently-labelled sequencing primers, and reaction buffer. Reaction mixes and templates are combined in the wells of an aluminum 96-well thermocycling plate. Thirty consecutive cycles of linear amplification (*i.e.*., one primer synthesis) steps are performed including denaturation, annealing of primer and template, and extension; *i.e.*, DNA synthesis. A heated lid with rubber gaskets on the thermocycling plate prevents evaporation without the need for an oil overlay.

Two sequencing protocols are used: one for dye-labelled primers and a second for dye-labelled dideoxy chain terminators. The shotgun sequencing involves use of four dye-labelled sequencing primers, one for each of the four terminator nucleotide. Each dye-primer is labelled with a different fluorescent dye, permitting the four individual reactions to be combined into one lane of the 373 DNA Sequencer for electrophoresis, detection, and base-calling. ABI currently supplies pre-mixed reaction mixes in bulk packages containing all the necessary non-template reagents for sequencing. Sequencing can be done with both plasmid and PCR- generated templates with both dye-primers and dye- terminators with approximately equal fidelity, although plasmid templates generally give longer usable sequences.

Thirty-two reactions are loaded per AB373 Sequencer each day, for a total of 960 samples. Electrophoresis is run overnight following the manufacturer's protocols, and the data is collected for twelve hours. Following electrophoresis and fluorescence detection, the ABI 373 performs automatic lane tracking and base-calling. The lane-tracking is confirmed visually. Each sequence electropherogram (or fluorescence lane trace) is inspected visually and assessed for quality. Trailing sequences of low quality are removed and the sequence itself is loaded via software to a Sybase database (archived daily to 8mm tape). Leading vector polylinker sequence is removed automatically by a software program. Average edited lengths of sequences from the standard ABI 373 are around 400 bp and depend mostly on the quality of the template used for the sequencing reaction. ABI 373 Sequencers converted to Stretch Liners provide a longer electrophoresis path prior to fluorescence detection and increase the average number of usable bases to 500-600 bp.

INFORMATICS

1. Data Management

A number of information management systems for a large-scale sequencing lab have been developed. (For review see, for instance, Kerlavage *et al.*, *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences*, IEEE Computer Society Press, Washington D. C., 585 (1993)) The system used to collect and assemble the sequence data was developed using the Sybase relational database management system and was designed to automate data flow wherever possible and to reduce user error. The database stores and correlates all information collected during the entire operation from template preparation to final analysis of the genome. Because the raw output of the ABI 373 Sequencers was based on a Macintosh platform and the data management system chosen was based on a Unix platform, it was necessary to design and implement a variety of multi- user, client-server applications which allow the raw data as well as analysis results to flow seamlessly into the database with a minimum of user effort.

2. Assembly

An assembly engine (TIGR Assembler) developed for the rapid and accurate assembly of thousands of sequence fragments was employed to generate contigs. The TIGR assembler simultaneously clusters and assembles fragments of the genome. In order to obtain the speed necessary to assemble more than 104 fragments, the algorithm builds a hash table of 12 bp oligonucleotide subsequences to generate a list of potential sequence fragment overlaps. The number of potential overlaps for each fragment determines which fragments are likely to fall into repetitive elements. Beginning with a single seed sequence fragment, TIGR Assembler extends the current contig by attempting to add the best matching fragment based on oligonucleotide content. The contig and candidate fragment are aligned using a modified version of the Smith-Waterman algorithm which provides for optimal gapped alignments (Waterman, M. S., *Methods*

*in Enzymology 164:*765 (1988)). The contig is extended by the fragment only if strict criteria for the quality of the match are met. The match criteria include the minimum length of overlap, the maximum length of an unmatched end, and the minimum percentage match. These criteria are automatically lowered by the algorithm in regions of minimal coverage and raised in regions with a possible repetitive element. The number of potential overlaps for each fragment determines which fragments are likely to fall into repetitive elements. Fragments representing the boundaries of repetitive elements and potentially chimeric fragments are often rejected based on partial mismatches at the ends of alignments and excluded from the current contig. TIGR Assembler is designed to take advantage of clone size information coupled with sequencing from both ends of each template. It enforces the constraint that sequence fragments from two ends of the same template point toward one another in the contig and are located within a certain range of base pairs (definable for each clone based on the known clone size range for a given library).

The process resulted in 744 contigs as represented by SEQ ID NOs:1-744.

3. Identifying Genes

The predicted coding regions of the *T. pallidum* genome were initially defined with the program GeneMark, which finds ORFs using a probabilistic classification technique. The predicted coding region sequences were used in searches against a database of all nucleotide sequences from GenBank (June, 1997), using the BLASTN search method to identify overlaps of 50 or more nucleotides with at least a 95% identity. Those ORFs with nucleotide sequence matches are shown in Table 1. The ORFs without such matches were translated to protein sequences and compared to a non-redundant database of known proteins generated by combining the Swiss-prot, PIR and GenPept databases. ORFs that matched a database protein with BLASTP probability less than or equal to 0.01 are shown in Table 2. The table also lists assigned functions based on the closest match in the databases. ORFs that did not match protein or nucleotide sequences in the databases at these levels are shown in Table 3.

ILLUSTRATIVE APPLICATIONS

1. Production of an Antibody to a *T. pallidum* Protein

Substantially pure protein or polypeptide is isolated from the transfected or transformed cells using any one of the methods known in the art. The protein can also be produced in a recombinant prokaryotic expression system, such as *E. coli*, or can be chemically synthesized. Concentration of protein in the final preparation is adjusted, for example, by concentration on an Amicon filter device, to the level of a few micrograms/ml. Monoclonal or polyclonal antibody to the protein can then be prepared as follows.

2. Monoclonal Antibody Production by Hybridoma Fusion

Monoclonal antibody to epitopes of any of the peptides identified and isolated as described can be prepared from murine hybridomas according to the classical method of Kohler,

G. and Milstein, C., *Nature 256*:495 (1975) or modifications of the methods thereof. Briefly, a mouse is repetitively inoculated with a few micrograms of the selected protein over a period of a few weeks. The mouse is then sacrificed, and the antibody producing cells of the spleen isolated. The spleen cells are fused by means of polyethylene glycol with mouse myeloma cells, and the excess unfused cells destroyed by growth of the system on selective media comprising aminopterin (HAT media). The successfully fused cells are diluted and aliquots of the dilution placed in wells of a microtiter plate where growth of the culture is continued. Antibody-producing clones are identified by detection of antibody in the supernatant fluid of the wells by immunoassay procedures, such as ELISA, as originally described by Engvall, E., *Meth. Enzymol. 70:*419 (1980), and modified methods thereof. Selected positive clones can be expanded and their monoclonal antibody product harvested for use. Detailed procedures for monoclonal antibody production are described in Davis, L. *et al., Basic Methods in Molecular Biology,* Elsevier, New York. Section 21-2 (1989).

3. Polyclonal Antibody Production by Immunization

Polyclonal antiserum containing antibodies to heterogenous epitopes of a single protein can be prepared by immunizing suitable animals with the expressed protein described above, which can be unmodified or modified to enhance immunogenicity. Effective polyclonal antibody production is affected by many factors related both to the antigen and the host species. For example, small molecules tend to be less immunogenic than others and may require the use of carriers and adjuvant. Also, host animals vary in response to site of inoculations and dose, with both inadequate or excessive doses of antigen resulting in low titer antisera. Small doses (ng level) of antigen administered at multiple intradermal sites appears to be most reliable. An effective immunization protocol for rabbits can be found in Vaitukaitis, J. *et al., J. Clin.* Endocrinol. Metab. 33:988-991 (1971).

Booster injections can be given at regular intervals, and antiserum harvested when antibody titer thereof, as determined semi-quantitatively, for example, by double immunodiffusion in agar against known concentrations of the antigen, begins to fall. See, for example, Ouchterlony, O. *et al.*, Chap. 19 in: *Handbook of Experimental Immunology,* Wier, D., ed, Blackwell (1973). Plateau concentration of antibody is usually in the range of 0.1 to 0.2 mg/ml of serum (about 12M). Affinity of the antisera for the antigen is determined by preparing competitive binding curves, as described, for example, by Fisher, D., Chap. 42 in: *Manual of Clinical Immunology,* second edition, Rose and Friedman, eds., Amer. Soc. For Microbiology, Washington, D. C. (1980)

Antibody preparations prepared according to either protocol are useful in quantitative immunoassays which determine concentrations of antigen-bearing substances in biological samples; they are also used semi- quantitatively or qualitatively to identify the presence of antigen in a biological sample. In addition, antibodies are useful in various animal models of pneumococcal disease as a means of evaluating the protein used to make the antibody as a

potential vaccine target or as a means of evaluating the antibody as a potential immunotherapeutic or immunoprophylactic reagent.

4. Preparation of PCR Primers and Amplification of DNA

Various fragments of the *T. pallidum* genome, such as those of Tables 1-3 and SEQ ID NOS: 1-744 can be used, in accordance with the present invention, to prepare PCR primers for a variety of uses. The PCR primers are preferably at least 15 bases, and more preferably at least 18 bases in length. When selecting a primer sequence, it is preferred that the primer pairs have approximately the same G/C ratio, so that melting temperatures are approximately the same. The PCR primers and amplified DNA of this Example find use in the Examples that follow.

5. Isolation of a Selected DNA Clone From T. pallidum

Three approaches are used to isolate a *T. pallidum* clone comprising a polynucleotide of the present invention from any *T. pallidum* genomic DNA library. The *T. pallidum* strain B31PU has been deposited as a convienent source for obtaining a *T. pallidum* strain although a wide varity of strains *T. pallidum* strains can be used which are known in the art.

T. pallidum genomic DNA is prepared using the following method. A 20ml overnight bacterial culture grown in a rich medium (e.g., Trypticase Soy Broth, Brain Heart Infusion broth or Super broth), pelleted, ished two times with TES (30mM Tris-pH 8.0, 25mM EDTA, 50mM NaCl), and resuspended in 5ml high salt TES (2.5M NaCl). Lysostaphin is added to final concentration of approx 50ug/ml and the mixture is rotated slowly 1 hour at 37C to make protoplast cells. The solution is then placed in incubator (or place in a shaking water bath) and warmed to 55C. Five hundred micro liter of 20% sarcosyl in TES (final concentration 2%) is then added to lyse the cells. Next, guanidine HCl is added to a final concentration of 7M (3.69g in 5.5 ml). The mixture is swirled slowly at 55C for 60-90 min (solution should clear). A CsCl gradient is then set up in SW41 ultra clear tubes using 2.0ml 5.7M CsCl and overlaying with 2.85M CsCl. The gradient is carefully overlayed with the DNA-containing GuHCl solution. The gradient is spun at 30,000 rpm, 20C for 24 hr and the lower DNA band is collected. The volume is increased to 5 ml with TE buffer. The DNA is then treated with protease K (10 ug/ml) overnight at 37 C, and precipitated with ethanol. The precipitated DNA is resuspended in a desired buffer.

In the first method, a plasmid is directly isolated by screening a plasmid *T. pallidum* genomic DNA library using a polynucleotide probe corresponding to a polynucleotide of the present invention. Particularly, a specific polynucleotide with 30-40 nucleotides is synthesized using an Applied Biosystems DNA synthesizer according to the sequence reported. The oligonucleotide is labeled, for instance, with ^{32}P-γ-ATP using T4 polynucleotide kinase and purified according to routine methods. (*See, e.g.*, Maniatis et al., Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Press, Cold Spring, NY (1982).) The library is

transformed into a suitable host, as indicated above (such as XL-1 Blue (Stratagene)) using techniques known to those of skill in the art. *See, e.g.*, Sambrook et al. MOLECULAR CLONING: A LABORATORY MANUAL (Cold Spring Harbor, N.Y. 2nd ed. 1989); Ausubel et al., CURRENT PROTOCALS IN MOLECULAR BIOLOGY (John Wiley and Sons, N.Y. 1989). The transformants are plated on 1.5% agar plates (containing the appropriate selection agent, e.g., ampicillin) to a density of about 150 transformants (colonies) per plate. These plates are screened using Nylon membranes according to routine methods for bacterial colony screening. *See, e.g.*, Sambrook et al. MOLECULAR CLONING: A LABORATORY MANUAL (Cold Spring Harbor, N.Y. 2nd ed. 1989); Ausubel et al., CURRENT PROTOCALS IN MOLECULAR BIOLOGY (John Wiley and Sons, N.Y. 1989) or other techniques known to those of skill in the art.

Alternatively, two primers of 15-25 nucleotides derived from the 5' and 3' ends of a polynucleotide of SEQ ID NOS:1-744 are synthesized and used to amplify the desired DNA by PCR using a *T. pallidum* genomic DNA prep as a template. PCR is carried out under routine conditions, for instance, in 25 µl of reaction mixture with 0.5 ug of the above DNA template. A convenient reaction mixture is 1.5-5 mM $MgCl_2$, 0.01% (w/v) gelatin, 20 µM each of dATP, dCTP, dGTP, dTTP, 25 pmol of each primer and 0.25 Unit of Taq polymerase. Thirty five cycles of PCR (denaturation at 94°C for 1 min; annealing at 55°C for 1 min; elongation at 72°C for 1 min) are performed with a Perkin-Elmer Cetus automated thermal cycler. The amplified product is analyzed by agarose gel electrophoresis and the DNA band with expected molecular weight is excised and purified. The PCR product is verified to be the selected sequence by subcloning and sequencing the DNA product.

Finally, overlapping oligos of the DNA sequences of SEQ ID NOS:1-744 can be chemically synthesized and used to generate a nucleotide sequence of desired length using PCR methods known in the art.

6(a). Expression and Purification Borrelia polypeptides in E. coli

The bacterial expression vector pQE60 is used for bacterial expression of some of the polypeptide fragements of the present invention. (QIAGEN, Inc., 9259 Eton Avenue, Chatsworth, CA, 91311). pQE60 encodes ampicillin antibiotic resistance ("Ampr") and contains a bacterial origin of replication ("ori"), an IPTG inducible promoter, a ribosome binding site ("RBS"), six codons encoding histidine residues that allow affinity purification using nickel-nitrilo-tri-acetic acid ("Ni-NTA") affinity resin (QIAGEN, Inc., *supra*) and suitable single restriction enzyme cleavage sites. These elements are arranged such that an inserted DNA fragment encoding a polypeptide expresses that polypeptide with the six His residues (i.e., a "6 X His tag") covalently linked to the carboxyl terminus of that polypeptide.

The DNA sequence encoding the desired portion of a *T. pallidum* protein of the present invention is amplified from *T. pallidum* genomic DNA using PCR oligonucleotide primers

which anneal to the 5' and 3' sequences coding for the portions of the *T. pallidum* polynucleotide shown in SEQ ID NOS:1-744. Additional nucleotides containing restriction sites to facilitate cloning in the pQE60 vector are added to the 5' and 3' sequences, respectively.

For cloning the mature protein, the 5' primer has a sequence containing an appropriate restriction site followed by nucleotides of the amino terminal coding sequence of the desired *T. pallidum* polynucleotide sequence in SEQ ID NOS:1-744. One of ordinary skill in the art would appreciate that the point in the protein coding sequence where the 5' and 3' primers begin may be varied to amplify a DNA segment encoding any desired portion of the complete protein shorter or longer than the mature form. The 3' primer has a sequence containing an appropriate restriction site followed by nucleotides complementary to the 3' end of the polypeptide coding sequence of SEQ ID NOS:1-744, excluding a stop codon, with the coding sequence aligned with the restriction site so as to maintain its reading frame with that of the six His codons in the pQE60 vector.

The amplified *T. pallidum* DNA fragment and the vector pQE60 are digested with restriction enzymes which recognize the sites in the primers and the digested DNAs are then ligated together. The *T. pallidum* DNA is inserted into the restricted pQE60 vector in a manner which places the *T. pallidum* protein coding region downstream from the IPTG-inducible promoter and in-frame with an initiating AUG and the six histidine codons.

The ligation mixture is transformed into competent *E. coli* cells using standard procedures such as those described by Sambrook et al., *supra*.. *E. coli* strain M15/rep4, containing multiple copies of the plasmid pREP4, which expresses the lac repressor and confers kanamycin resistance ("Kanr"), is used in carrying out the illustrative example described herein. This strain, which is only one of many that are suitable for expressing a *T. pallidum* polypeptide, is available commercially (QIAGEN, Inc., *supra*). Transformants are identified by their ability to grow on LB agar plates in the presence of ampicillin and kanamycin. Plasmid DNA is isolated from resistant colonies and the identity of the cloned DNA confirmed by restriction analysis, PCR and DNA sequencing.

Clones containing the desired constructs are grown overnight ("O/N") in liquid culture in LB media supplemented with both ampicillin (100 μg/ml) and kanamycin (25 μg/ml). The O/N culture is used to inoculate a large culture, at a dilution of approximately 1:25 to 1:250. The cells are grown to an optical density at 600 nm ("OD600") of between 0.4 and 0.6. Isopropyl-β-D-thiogalactopyranoside ("IPTG") is then added to a final concentration of 1 mM to induce transcription from the lac repressor sensitive promoter, by inactivating the lacI repressor. Cells subsequently are incubated further for 3 to 4 hours. Cells then are harvested by centrifugation.

The cells are then stirred for 3-4 hours at 4°C in 6M guanidine-HCl, pH 8. The cell debris is removed by centrifugation, and the supernatant containing the *T. pallidum* polypeptide is loaded onto a nickel-nitrilo-tri-acetic acid ("Ni-NTA") affinity resin column (QIAGEN, Inc., *supra*). Proteins with a 6 x His tag bind to the Ni-NTA resin with high affinity are purified in a

simple one-step procedure (for details see: The QIAexpressionist, 1995, QIAGEN, Inc., *supra*). Briefly the supernatant is loaded onto the column in 6 M guanidine-HCl, pH 8, the column is first washed with 10 volumes of 6 M guanidine-HCl, pH 8, then washed with 10 volumes of 6 M guanidine-HCl pH 6, and finally the *T. pallidum* polypeptide is eluted with 6 M guanidine-HCl, pH 5.

The purified protein is then renatured by dialyzing it against phosphate-buffered saline (PBS) or 50 mM Na-acetate, pH 6 buffer plus 200 mM NaCl. Alternatively, the protein could be successfully refolded while immobilized on the Ni-NTA column. The recommended conditions are as follows: renature using a linear 6M-1M urea gradient in 500 mM NaCl, 20% glycerol, 20 mM Tris/HCl pH 7.4, containing protease inhibitors. The renaturation should be performed over a period of 1.5 hours or more. After renaturation the proteins can be eluted by the addition of 250 mM immidazole. Immidazole is removed by a final dialyzing step against PBS or 50 mM sodium acetate pH 6 buffer plus 200 mM NaCl. The purified protein is stored at 4° C or frozen at -80° C.

The polypeptide of the present invention are also prepared using a non-denaturing protein purification method. For these polypeptides, the cell pellet from each liter of culture is resuspended in 25 mls of Lysis Buffer A at 4°C (Lysis Buffer A = 50 mM Na-phosphate, 300 mM NaCl, 10 mM 2-mercaptoethanol, 10% Glycerol, pH 7.5 with 1 tablet of Complete EDTA-free protease inhibitor cocktail (Boehringer Mannheim #1873580) per 50 ml of buffer). Absorbance at 550 nm is approximately 10-20 O.D./ml. The suspension is then put through three freeze/thaw cycles from -70°C (using a ethanol-dry ice bath) up to room temperature. The cells are lysed via sonication in short 10 sec bursts over 3 minutes at approximately 80W while kept on ice. The sonicated sample is then centrifuged at 15,000 RPM for 30 minutes at 4°C. The supernatant is passed through a column containing 1.0 ml of CL-4B resin to pre-clear the sample of any proteins that may bind to agarose non-specifically, and the flow-through fraction is collected.

The pre-cleared flow-through is applied to a nickel-nitrilo-tri-acetic acid ("Ni-NTA") affinity resin column (Quiagen, Inc., *supra*). Proteins with a 6 X His tag bind to the Ni-NTA resin with high affinity and can be purified in a simple one-step procedure. Briefly, the supernatant is loaded onto the column in Lysis Buffer A at 4°C, the column is first washed with 10 volumes of Lysis Buffer A until the A280 of the eluate returns to the baseline. Then, the column is washed with 5 volumes of 40 mM Imidazole (92% Lysis Buffer A / 8% Buffer B) (Buffer B = 50 mM Na-Phosphate, 300 mM NaCl, 10% Glycerol, 10 mM 2-mercaptoethanol, 500 mM Imidazole, pH of the final buffer should be 7.5). The protein is eluted off of the column with a series of increasing Imidazole solutions made by adjusting the ratios of Lysis Buffer A to Buffer B. Three different concentrations are used: 3 volumes of 75 mM Imidazole, 3 volumes of 150 mM Imidazole, 5 volumes of 500 mM Imidazole. The fractions containing the purified protein are analyzed using 8 %, 10 % or 14% SDS-PAGE depending on the protein size. The purified protein is then dialyzed 2X against phosphate-buffered saline (PBS) in order to place it

into an easily workable buffer. The purified protein is stored at 4° C or frozen at -80°.

The following alternative method may be used to purify *T. pallidum* expressed in *E coli* when it is present in the form of inclusion bodies. Unless otherwise specified, all of the following steps are conducted at 4-10°C.

Upon completion of the production phase of the *E. coli* fermentation, the cell culture is cooled to 4-10°C and the cells are harvested by continuous centrifugation at 15,000 rpm (Heraeus Sepatech). On the basis of the expected yield of protein per unit weight of cell paste and the amount of purified protein required, an appropriate amount of cell paste, by weight, is suspended in a buffer solution containing 100 mM Tris, 50 mM EDTA, pH 7.4. The cells are dispersed to a homogeneous suspension using a high shear mixer.

The cells are then lysed by passing the solution through a microfluidizer (Microfuidics, Corp. or APV Gaulin, Inc.) twice at 4000-6000 psi. The homogenate is then mixed with NaCl solution to a final concentration of 0.5 M NaCl, followed by centrifugation at 7000 x *g* for 15 min. The resultant pellet is washed again using 0.5M NaCl, 100 mM Tris, 50 mM EDTA, pH 7.4.

The resulting washed inclusion bodies are solubilized with 1.5 M guanidine hydrochloride (GuHCl) for 2-4 hours. After 7000 x *g* centrifugation for 15 min., the pellet is discarded and the *T. pallidum* polypeptide-containing supernatant is incubated at 4°C overnight to allow further GuHCl extraction.

Following high speed centrifugation (30,000 x *g*) to remove insoluble particles, the GuHCl solubilized protein is refolded by quickly mixing the GuHCl extract with 20 volumes of buffer containing 50 mM sodium, pH 4.5, 150 mM NaCl, 2 mM EDTA by vigorous stirring. The refolded diluted protein solution is kept at 4°C without mixing for 12 hours prior to further purification steps.

To clarify the refolded *T. pallidum* polypeptide solution, a previously prepared tangential filtration unit equipped with 0.16 µm membrane filter with appropriate surface area (e.g., Filtron), equilibrated with 40 mM sodium acetate, pH 6.0 is employed. The filtered sample is loaded onto a cation exchange resin (e.g., Poros HS-50, Perseptive Biosystems). The column is washed with 40 mM sodium acetate, pH 6.0 and eluted with 250 mM, 500 mM, 1000 mM, and 1500 mM NaCl in the same buffer, in a stepwise manner. The absorbance at 280 mm of the effluent is continuously monitored. Fractions are collected and further analyzed by SDS-PAGE.

Fractions containing the *T. pallidum* polypeptide are then pooled and mixed with 4 volumes of water. The diluted sample is then loaded onto a previously prepared set of tandem columns of strong anion (Poros HQ-50, Perseptive Biosystems) and weak anion (Poros CM-20, Perseptive Biosystems) exchange resins. The columns are equilibrated with 40 mM sodium acetate, pH 6.0. Both columns are washed with 40 mM sodium acetate, pH 6.0, 200 mM NaCl.

The CM-20 column is then eluted using a 10 column volume linear gradient ranging from 0.2 M NaCl, 50 mM sodium acetate, pH 6.0 to 1.0 M NaCl, 50 mM sodium acetate, pH 6.5. Fractions are collected under constant A_{280} monitoring of the effluent. Fractions containing the *T. pallidum* polypeptide (determined, for instance, by 16% SDS-PAGE) are then pooled.

The resultant *T. pallidum* polypeptide exhibits greater than 95% purity after the above refolding and purification steps. No major contaminant bands are observed from Commassie blue stained 16% SDS-PAGE gel when 5 μg of purified protein is loaded. The purified protein is also tested for endotoxin/LPS contamination, and typically the LPS content is less than 0.1 ng/ml according to LAL assays.

6(b). Alternative Expression and Purification Borrelia polypeptides in E. coli

Tthe vector pQE10 is alternatively used to clone and express some of the polypeptides of the present invention for use in the soft tissue and systemic infection models discussed below. The difference being such that an inserted DNA fragment encoding a polypeptide expresses that polypeptide with the six His residues (i.e., a "6 X His tag") covalently linked to the amino terminus of that polypeptide. The bacterial expression vector pQE10 (QIAGEN, Inc., 9259 Eton Avenue, Chatsworth, CA, 91311) was used in this example . The components of the pQE10 plasmid are arranged such that the inserted DNA sequence encoding a polypeptide of the present invention expresses the polypeptide with the six His residues (*i.e.*, a "6 X His tag")) covalently linked to the amino terminus.

The DNA sequences encoding the desired portions of a polypeptide of SEQ ID NOS:1-744 were amplified using PCR oligonucleotide primers from genomic *T. pallidum* DNA. The PCR primers anneal to the nucleotide sequences encoding the desired amino acid sequence of a polypeptide of the present invention. Additional nucleotides containing restriction sites to facilitate cloning in the pQE10 vector were added to the 5' and 3' primer sequences, respectively.

For cloning a polypeptide of the present invention, the 5' and 3' primers were selected to amplify their respective nucleotide coding sequences. One of ordinary skill in the art would appreciate that the point in the protein coding sequence where the 5' and 3' primers begins may be varied to amplify a DNA segment encoding any desired portion of a polypeptide of the present invention. The 5' primer was designed so the coding sequence of the 6 X His tag is aligned with the restriction site so as to maintain its reading frame with that of *T. pallidum* polypeptide. The 3' was designed to include an stop codon. The amplified DNA fragment was then cloned, and the protein expressed, as described above for the pQE60 plasmid.

The DNA sequences encoding the amino acid sequences of SEQ ID NOS:1-744 may also be cloned and expressed as fusion proteins by a protocol similar to that described directly above, wherein the pET-32b(+) vector (Novagen, 601 Science Drive, Madison, WI 53711) is preferentially used in place of pQE10.

The above methods are not limited to the polypeptide fragements actually produced. The above method, like the methods below, can be used to produce either full length polypeptides or desired fragements therof.

6(c). Alternative Expression and Purification of Borrelia polypeptides in E. coli

The bacterial expression vector pQE60 is used for bacterial expression in this example (QIAGEN, Inc., 9259 Eton Avenue, Chatsworth, CA, 91311). However, in this example, the polypeptide coding sequence is inserted such that translation of the six His codons is prevented and, therefore, the polypeptide is produced with no 6 X His tag.

The DNA sequence encoding the desired portion of the *T. pallidum* amino acid sequence is amplified from an *T. pallidum* genomic DNA prep the deposited DNA clones using PCR oligonucleotide primers which anneal to the 5' and 3' nucleotide sequences corresponding to the desired portion of the *T. pallidum* polypeptides. Additional nucleotides containing restriction sites to facilitate cloning in the pQE60 vector are added to the 5' and 3' primer sequences.

For cloning a *T. pallidum* polypeptides of the present invention, 5' and 3' primers are selected to amplify their respective nucleotide coding sequences. One of ordinary skill in the art would appreciate that the point in the protein coding sequence where the 5' and 3' primers begin may be varied to amplify a DNA segment encoding any desired portion of a polypeptide of the present invention. The 3' and 5' primers contain appropriate restriction sites followed by nucleotides complementary to the 5' and 3' ends of the coding sequence respectively. The 3' primer is additionally designed to include an in-frame stop codon.

The amplified *T. pallidum* DNA fragments and the vector pQE60 are digested with restriction enzymes recognizing the sites in the primers and the digested DNAs are then ligated together. Insertion of the *T. pallidum* DNA into the restricted pQE60 vector places the *T. pallidum* protein coding region including its associated stop codon downstream from the IPTG-inducible promoter and in-frame with an initiating AUG. The associated stop codon prevents translation of the six histidine codons downstream of the insertion point.

The ligation mixture is transformed into competent *E. coli* cells using standard procedures such as those described by Sambrook et al. *E. coli* strain M15/rep4, containing multiple copies of the plasmid pREP4, which expresses the lac repressor and confers kanamycin resistance ("Kanr"), is used in carrying out the illustrative example described herein. This strain, which is only one of many that are suitable for expressing *T. pallidum* polypeptide, is available commercially (QIAGEN, Inc., *supra*). Transformants are identified by their ability to grow on LB plates in the presence of ampicillin and kanamycin. Plasmid DNA is isolated from resistant colonies and the identity of the cloned DNA confirmed by restriction analysis, PCR and DNA sequencing.

Clones containing the desired constructs are grown overnight ("O/N") in liquid culture in LB media supplemented with both ampicillin (100 µg/ml) and kanamycin (25 µg/ml). The O/N

culture is used to inoculate a large culture, at a dilution of approximately 1:25 to 1:250. The cells are grown to an optical density at 600 nm ("OD600") of between 0.4 and 0.6. isopropyl-b-D-thiogalactopyranoside ("IPTG") is then added to a final concentration of 1 mM to induce transcription from the *lac* repressor sensitive promoter, by inactivating the lacI repressor. Cells subsequently are incubated further for 3 to 4 hours. Cells then are harvested by centrifugation.

To purify the *T. pallidum* polypeptide, the cells are then stirred for 3-4 hours at 4°C in 6M guanidine-HCl, pH 8. The cell debris is removed by centrifugation, and the supernatant containing the *T. pallidum* polypeptide is dialyzed against 50 mM Na-acetate buffer pH 6, supplemented with 200 mM NaCl. Alternatively, the protein can be successfully refolded by dialyzing it against 500 mM NaCl, 20% glycerol, 25 mM Tris/HCl pH 7.4, containing protease inhibitors. After renaturation the protein can be purified by ion exchange, hydrophobic interaction and size exclusion chromatography. Alternatively, an affinity chromatography step such as an antibody column can be used to obtain pure *T. pallidum* polypeptide. The purified protein is stored at 4° C or frozen at -80° C.

The following alternative method may be used to purify *T. pallidum* polypeptides expressed in *E coli* when it is present in the form of inclusion bodies. Unless otherwise specified, all of the following steps are conducted at 4-10°C.

Upon completion of the production phase of the *E. coli* fermentation, the cell culture is cooled to 4-10°C and the cells are harvested by continuous centrifugation at 15,000 rpm (Heraeus Sepatech). On the basis of the expected yield of protein per unit weight of cell paste and the amount of purified protein required, an appropriate amount of cell paste, by weight, is suspended in a buffer solution containing 100 mM Tris, 50 mM EDTA, pH 7.4. The cells are dispersed to a homogeneous suspension using a high shear mixer.

The cells ware then lysed by passing the solution through a microfluidizer (Microfuidics, Corp. or APV Gaulin, Inc.) twice at 4000-6000 psi. The homogenate is then mixed with NaCl solution to a final concentration of 0.5 M NaCl, followed by centrifugation at 7000 x *g* for 15 min. The resultant pellet is washed again using 0.5M NaCl, 100 mM Tris, 50 mM EDTA, pH 7.4.

The resulting washed inclusion bodies are solubilized with 1.5 M guanidine hydrochloride (GuHCl) for 2-4 hours. After 7000 x *g* centrifugation for 15 min., the pellet is discarded and the *T. pallidum* polypeptide-containing supernatant is incubated at 4°C overnight to allow further GuHCl extraction.

Following high speed centrifugation (30,000 x *g*) to remove insoluble particles, the GuHCl solubilized protein is refolded by quickly mixing the GuHCl extract with 20 volumes of buffer containing 50 mM sodium, pH 4.5, 150 mM NaCl, 2 mM EDTA by vigorous stirring. The refolded diluted protein solution is kept at 4°C without mixing for 12 hours prior to further

purification steps.

To clarify the refolded *T. pallidum* polypeptide solution, a previously prepared tangential filtration unit equipped with 0.16 μm membrane filter with appropriate surface area (e.g., Filtron), equilibrated with 40 mM sodium acetate, pH 6.0 is employed. The filtered sample is loaded onto a cation exchange resin (e.g., Poros HS-50, Perseptive Biosystems). The column is washed with 40 mM sodium acetate, pH 6.0 and eluted with 250 mM, 500 mM, 1000 mM, and 1500 mM NaCl in the same buffer, in a stepwise manner. The absorbance at 280 mm of the effluent is continuously monitored. Fractions are collected and further analyzed by SDS-PAGE.

Fractions containing the *T. pallidum* polypeptide are then pooled and mixed with 4 volumes of water. The diluted sample is then loaded onto a previously prepared set of tandem columns of strong anion (Poros HQ-50, Perseptive Biosystems) and weak anion (Poros CM-20, Perseptive Biosystems) exchange resins. The columns are equilibrated with 40 mM sodium acetate, pH 6.0. Both columns are washed with 40 mM sodium acetate, pH 6.0, 200 mM NaCl. The CM-20 column is then eluted using a 10 column volume linear gradient ranging from 0.2 M NaCl, 50 mM sodium acetate, pH 6.0 to 1.0 M NaCl, 50 mM sodium acetate, pH 6.5. Fractions are collected under constant A_{280} monitoring of the effluent. Fractions containing the *T. pallidum* polypeptide (determined, for instance, by 16% SDS-PAGE) are then pooled.

The resultant *T. pallidum* polypeptide exhibits greater than 95% purity after the above refolding and purification steps. No major contaminant bands are observed from Commassie blue stained 16% SDS-PAGE gel when 5 μg of purified protein is loaded. The purified protein is also tested for endotoxin/LPS contamination, and typically the LPS content is less than 0.1 ng/ml according to LAL assays.

6(d). Cloning and Expression of T. pallidum in Other Bacteria

T. pallidum polypeptides can also be produced in: *T. pallidum* using the methods of S. Skinner et al., (1988) Mol. Microbiol. 2:289-297 or J. I. Moreno (1996) Protein Expr. Purif. 8(3):332-340; *Lactobacillus* using the methods of C. Rush et al., 1997 Appl. Microbiol. Biotechnol. 47(5):537-542; or in *Bacillus subtilis* using the methods Chang et al., U.S. Patent No. 4,952,508.

7. Cloning and Expression in COS Cells

A *T. pallidum* expression plasmid is made by cloning a portion of the DNA encoding a *T. pallidum* polypeptide into the expression vector pDNAI/Amp or pDNAIII (which can be obtained from Invitrogen, Inc.). The expression vector pDNAI/amp contains: (1) an *E. coli* origin of replication effective for propagation in *E. coli* and other prokaryotic cells; (2) an ampicillin resistance gene for selection of plasmid-containing prokaryotic cells; (3) an SV40 origin of replication for propagation in eukaryotic cells; (4) a CMV promoter, a polylinker, an SV40 intron; (5) several codons encoding a hemagglutinin fragment (i.e., an "HA" tag to

facilitate purification) followed by a termination codon and polyadenylation signal arranged so that a DNA can be conveniently placed under expression control of the CMV promoter and operably linked to the SV40 intron and the polyadenylation signal by means of restriction sites in the polylinker. The HA tag corresponds to an epitope derived from the influenza hemagglutinin protein described by Wilson et al. 1984 Cell 37:767. The fusion of the HA tag to the target protein allows easy detection and recovery of the recombinant protein with an antibody that recognizes the HA epitope. pDNAIII contains, in addition, the selectable neomycin marker.

A DNA fragment encoding a *T. pallidum* polypeptide is cloned into the polylinker region of the vector so that recombinant protein expression is directed by the CMV promoter. The plasmid construction strategy is as follows. The DNA from a *T. pallidum* genomic DNA prep is amplified using primers that contain convenient restriction sites, much as described above for construction of vectors for expression of *T. pallidum* in *E. coli*. The 5' primer contains a Kozak sequence, an AUG start codon, and nucleotides of the 5' coding region of the *T. pallidum* polypeptide. The 3' primer, contains nucleotides complementary to the 3' coding sequence of the *T. pallidum* DNA, a stop codon, and a convenient restriction site.

The PCR amplified DNA fragment and the vector, pDNAI/Amp, are digested with appropriate restriction enzymes and then ligated. The ligation mixture is transformed into an appropriate *E. coli* strain such as SURE™ (Stratagene Cloning Systems, La Jolla, CA 92037), and the transformed culture is plated on ampicillin media plates which then are incubated to allow growth of ampicillin resistant colonies. Plasmid DNA is isolated from resistant colonies and examined by restriction analysis or other means for the presence of the fragment encoding the *T. pallidum* polypeptide

For expression of a recombinant *T. pallidum* polypeptide, COS cells are transfected with an expression vector, as described above, using DEAE-dextran, as described, for instance, by Sambrook et al. (*supra*). Cells are incubated under conditions for expression of *T. pallidum* by the vector.

Expression of the *T. pallidum*-HA fusion protein is detected by radiolabeling and immunoprecipitation, using methods described in, for example Harlow et al., *supra*.. To this end, two days after transfection, the cells are labeled by incubation in media containing ^{35}S-cysteine for 8 hours. The cells and the media are collected, and the cells are washed and the lysed with detergent-containing RIPA buffer: 150 mM NaCl, 1% NP-40, 0.1% SDS, 1% NP-40, 0.5% DOC, 50 mM TRIS, pH 7.5, as described by Wilson et al. (*supra*). Proteins are precipitated from the cell lysate and from the culture media using an HA-specific monoclonal antibody. The precipitated proteins then are analyzed by SDS-PAGE and autoradiography. An expression product of the expected size is seen in the cell lysate, which is not seen in negative controls.

8. Cloning and Expression in CHO Cells

The vector pC4 is used for the expression of *T. pallidum* polypeptide in this example. Plasmid pC4 is a derivative of the plasmid pSV2-dhfr (ATCC Accession No. 37146). The plasmid contains the mouse DHFR gene under control of the SV40 early promoter. Chinese hamster ovary cells or other cells lacking dihydrofolate activity that are transfected with these plasmids can be selected by growing the cells in a selective medium (alpha minus MEM, Life Technologies) supplemented with the chemotherapeutic agent methotrexate. The amplification of the DHFR genes in cells resistant to methotrexate (MTX) has been well documented. *See, e.g.*, Alt et al., 1978, J. Biol. Chem. 253:1357-1370; Hamlin et al., 1990, Biochem. et Biophys. Acta, 1097:107-143; Page et al., 1991, Biotechnology 9:64-68. Cells grown in increasing concentrations of MTX develop resistance to the drug by overproducing the target enzyme, DHFR, as a result of amplification of the DHFR gene. If a second gene is linked to the DHFR gene, it is usually co-amplified and over-expressed. It is known in the art that this approach may be used to develop cell lines carrying more than 1,000 copies of the amplified gene(s). Subsequently, when the methotrexate is withdrawn, cell lines are obtained which contain the amplified gene integrated into one or more chromosome(s) of the host cell.

Plasmid pC4 contains the strong promoter of the long terminal repeat (LTR) of the Rouse Sarcoma Virus, for expressing a polypeptide of interest, Cullen, et al. (1985) Mol. Cell. Biol. 5:438-447; plus a fragment isolated from the enhancer of the immediate early gene of human cytomegalovirus (CMV), Boshart, et al., 1985, Cell 41:521-530. Downstream of the promoter are the following single restriction enzyme cleavage sites that allow the integration of the genes: *Bam* HI, *Xba* I, and *Asp* 718. Behind these cloning sites the plasmid contains the 3' intron and polyadenylation site of the rat preproinsulin gene. Other high efficiency promoters can also be used for the expression, e.g., the human ß-actin promoter, the SV40 early or late promoters or the long terminal repeats from other retroviruses, e.g., HIV and HTLVI. Clontech's Tet-Off and Tet-On gene expression systems and similar systems can be used to express the *T. pallidum* polypeptide in a regulated way in mammalian cells (Gossen et al., 1992, Proc. Natl. Acad. Sci. USA 89:5547-5551. For the polyadenylation of the mRNA other signals, e.g., from the human growth hormone or globin genes can be used as well. Stable cell lines carrying a gene of interest integrated into the chromosomes can also be selected upon co-transfection with a selectable marker such as gpt, G418 or hygromycin. It is advantageous to use more than one selectable marker in the beginning, e.g., G418 plus methotrexate.

The plasmid pC4 is digested with the restriction enzymes and then dephosphorylated using calf intestinal phosphates by procedures known in the art. The vector is then isolated from a 1% agarose gel. The DNA sequence encoding the *T. pallidum* polypeptide is amplified using PCR oligonucleotide primers corresponding to the 5' and 3' sequences of the desired portion of the gene. A 5' primer containing a restriction site, a Kozak sequence, an AUG start codon, and nucleotides of the 5' coding region of the *T. pallidum* polypeptide is synthesized and used. A 3' primer, containing a restriction site, stop codon, and nucleotides complementary to the 3' coding sequence of the *T. pallidum* polypeptides is synthesized and used. The amplified fragment is

digested with the restriction endonucleases and then purified again on a 1% agarose gel. The isolated fragment and the dephosphorylated vector are then ligated with T4 DNA ligase. *E. coli* HB101 or XL-1 Blue cells are then transformed and bacteria are identified that contain the fragment inserted into plasmid pC4 using, for instance, restriction enzyme analysis.

Chinese hamster ovary cells lacking an active DHFR gene are used for transfection. Five μg of the expression plasmid pC4 is cotransfected with 0.5 μg of the plasmid pSVneo using a lipid-mediated transfection agent such as Lipofectin™ or LipofectAMINE.™ (LifeTechnologies Gaithersburg, MD). The plasmid pSV2-neo contains a dominant selectable marker, the *neo* gene from Tn5 encoding an enzyme that confers resistance to a group of antibiotics including G418. The cells are seeded in alpha minus MEM supplemented with 1 mg/ml G418. After 2 days, the cells are trypsinized and seeded in hybridoma cloning plates (Greiner, Germany) in alpha minus MEM supplemented with 10, 25, or 50 ng/ml of methotrexate plus 1 mg/ml G418. After about 10-14 days single clones are trypsinized and then seeded in 6-well petri dishes or 10 ml flasks using different concentrations of methotrexate (50 nM, 100 nM, 200 nM, 400 nM, 800 nM). Clones growing at the highest concentrations of methotrexate are then transferred to new 6-well plates containing even higher concentrations of methotrexate (1 μM, 2 μM, 5 μM, 10 mM, 20 mM). The same procedure is repeated until clones are obtained which grow at a concentration of 100-200 μM. Expression of the desired gene product is analyzed, for instance, by SDS-PAGE and Western blot or by reversed phase HPLC analysis.

The disclosure of all publications (including patents, patent applications, journal articles, laboratory manuals, books, or other documents) cited herein are hereby incorporated by reference in their entireties SEQ ID NOS: 1-744 are hereby incorporated into the specification by reference.

The present invention is not to be limited in scope by the specific embodiments described herein, which are intended as single illustrations of individual aspects of the invention. Functionally equivalent methods and components are within the scope of the invention, in addition to those shown and described herein and will become apparant to those skilled in the art from the foregoing description and accompanying drawings. Such modifications are intended to fall within the scope of the appended claims.

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

Contig ID	Orf ID	Start (nt)	Stop (nt)	match acession	match gene name	percent ident	HSP nt length
1	1	219	4	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds	93	202
1	2	110	493	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds	97	199
1	3	1167	226	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds	99	829
1	4	1237	1644	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds	98	285
3	1	225	2105	gb|U32683|	Treponema pallidum cytoplasmic filament protein A (cfpA) gene, complete cds	99	1800
16	44	25607	26131	emb|X61228| TPTP33G	T.pallidum Tp33 gene (partial)	96	84
18	1	647	1471	gb|U36839|	Treponema pallidum putative switch protein FliM (fliM) gene, partial cds, and FliY (fliY) gene, complete cds, export apparatus proteins FliP (fliP), FliQ (fliQ), FliR (fliR) and FlhB (flhB), signal transducing receptor FlhA (flhA), GTP binding protein >	100	797
18	2	1687	2574	gb|U36839|	Treponema pallidum putative switch protein FliM (fliM) gene, partial cds, and FliY (fliY) gene, complete cds, export apparatus proteins FliP (fliP), FliQ (fliQ), FliR (fliR) and FlhB (flhB), signal transducing receptor FlhA (flhA), GTP binding protein >	98	843
18	3	3200	2589	gb|U36839|	Treponema pallidum putative switch protein FliM (fliM) gene, partial cds, and FliY (fliY) gene, complete cds, export apparatus proteins FliP (fliP),	100	459

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

					FliQ (fliQ), FliR (fliR) and FlhB (flhB), signal transducing receptor FlhA (flhA), GTP binding protein >		
21	4	3095	2532	gb|M32401|	T.pallidum pallidum antigen TyF1 gene, complete cds	99	426
21	5	3151	3807	gb|M32401|	T.pallidum pallidum antigen TyF1 gene, complete cds	98	657
25	5	3042	2275	gb|U97363|	Treponema pallidum FlaA homolog-1 and FlaA homolog-2 genes, complete cds	98	269
26	15	8262	9125	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	68	417
26	16	10214	8373	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	68	417
26	17	9123	9719	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	85	122
26	19	11173	10226	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	99	916
26	20	11397	11095	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	97	228
26	21	12800	11316	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	99	1079
26	22	11568	11777	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	100	210
26	23	12824	13621	emb|X57836| TPTMBCL	T.pallidum tmbC gene for membrane lipoprotein	99	662
29	1	3	806	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	64	407
29	2	884	165	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	64	407
31	1	1	510	gb|U70661|	Treponema pallidum GTP-binding protein, mccF-like protein and ATP-dependent DNA helicase (RecG) genes, complete cds	100	102

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

32	1	3	458	gb|M26525|	T.pallidum endoflagellar sheath protein (flaA) gene, 3' end	100	423
32	2	358	810	gb|M26525|	T.pallidum endoflagellar sheath protein (flaA) gene, 3' end	100	384
34	7	5817	5134	gb|U32683|	Treponema pallidum cytoplasmic filament protein A (cfpA) gene, complete cds	98	65
35	20	12350	11691	gb|U61534|	Treponema pallidum 2-phospho-D-glycerate hydrolase (Eno) gene, complete cds	100	610
35	21	12918	12304	gb|U61534|	Treponema pallidum 2-phospho-D-glycerate hydrolase (Eno) gene, complete cds	99	604
35	22	13933	13091	gb|U93844|	Treponema pallidum lipoprotein homolog (tpN32) gene, complete cds, and 2-phospho-D-glycerate hydrolase (eno) gene, partial cds	100	780
35	23	14037	13792	gb|U93844|	Treponema pallidum lipoprotein homolog (tpN32) gene, complete cds, and 2-phospho-D-glycerate hydrolase (eno) gene, partial cds	99	169
35	24	14634	14050	gb|U93844|	Treponema pallidum lipoprotein homolog (tpN32) gene, complete cds, and 2-phospho-D-glycerate hydrolase (eno) gene, partial cds	100	544
35	25	15748	14915	gb|U93844|	Treponema pallidum lipoprotein homolog (tpN32) gene, complete cds, and 2-phospho-D-glycerate hydrolase (eno) gene, partial cds	100	834
35	26	15679	16056	gb|U97358|	Treponema pallidum 29K protein gene, complete cds	99	301
35	27	16719	15802	gb|U97358|	Treponema pallidum 29K protein gene, complete cds	99	178
36	1	526	1137	gb|U57756|	Treponema pallidum alanine racemase gene, partial cds	100	60
36	2	1157	2299	gb|U57756|	Treponema pallidum alanine racemase gene, partial cds	100	605
36	24	19592	20290	gb|U97363|	Treponema pallidum FlaA homolog-1 and FlaA homolog-2 genes, complete cds	98	261

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

36	26	20754	20308	gb|U97363|	Treponema pallidum FlaA homolog-1 and FlaA homolog-2 genes, complete cds	96	395
45	12	9409	10152	gb|M73825|	Treponema pallidum 1-pyrroline-5-carboxylate reductase gene, complete cds	100	291
46	1	547	101	gb|U97361|	Treponema pallidum H-ATPase homolog gene, partial cds	99	403
51	6	6397	7371	emb|X61226| TPTP75G	T.pallidum Tp75 gene (partial)	100	141
54	1	652	2	gb|U28219|	Treponema pallidum flagellar hook (flgE), (orf4), flagellar motor protein (motA), flagellar motor protein (motB), (fliL), and flagellar switch protein (fliM) genes, complete cds, and (fliY) gene, partial cds	99	565
54	2	1306	905	gb|U28219|	Treponema pallidum flagellar hook (flgE), (orf4), flagellar motor protein (motA), flagellar motor protein (motB), (fliL), and flagellar switch protein (fliM) genes, complete cds, and (fliY) gene, partial cds	99	388
54	3	1848	1237	gb|U28219|	Treponema pallidum flagellar hook (flgE), (orf4), flagellar motor protein (motA), flagellar motor protein (motB), (fliL), and flagellar switch protein (fliM) genes, complete cds, and (fliY) gene, partial cds	100	503
54	4	1243	1479	gb|U28219|	Treponema pallidum flagellar hook (flgE), (orf4), flagellar motor protein (motA), flagellar motor protein (motB), (fliL), and flagellar switch protein (fliM) genes, complete cds, and (fliY) gene, partial cds	100	194
54	5	2126	1755	gb|U28219|	Treponema pallidum flagellar hook (flgE), (orf4), flagellar motor protein (motA), flagellar motor protein (motB), (fliL), and flagellar switch protein (fliM) genes, complete cds, and (fliY) gene, partial cds	96	192

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

54	6	2329	1913	gb|U28219|	Treponema pallidum flagellar hook (flgE), (orf4), flagellar motor protein (motA), flagellar motor protein (motB), (fliL), and flagellar switch protein (fliM) genes, complete cds, and (fliY) gene, partial cds	98	380
54	7	2735	2364	gb|U42012|	Treponema pallidum putative treponemal aqueous protein Tap1 (tap1) gene, and flagellar hook assembly scaffolding protein (flgD) gene, complete cds	100	283
54	8	2872	2573	gb|U42012|	Treponema pallidum putative treponemal aqueous protein Tap1 (tap1) gene, and flagellar hook assembly scaffolding protein (flgD) gene, complete cds	100	226
54	9	3388	2900	gb|U42012|	Treponema pallidum putative treponemal aqueous protein Tap1 (tap1) gene, and flagellar hook assembly scaffolding protein (flgD) gene, complete cds	99	414
54	10	3169	4188	gb|U42012|	Treponema pallidum putative treponemal aqueous protein Tap1 (tap1) gene, and flagellar hook assembly scaffolding protein (flgD) gene, complete cds	99	448
54	11	4191	3238	gb|U42012|	Treponema pallidum putative treponemal aqueous protein Tap1 (tap1) gene, and flagellar hook assembly scaffolding protein (flgD) gene, complete cds	99	448
54	12	3683	3309	gb|U42012|	Treponema pallidum putative treponemal aqueous protein Tap1 (tap1) gene, and flagellar hook assembly scaffolding protein (flgD) gene, complete cds	99	140
54	13	4529	4014	gb|U42012|	Treponema pallidum putative treponemal aqueous protein Tap1 (tap1) gene, and flagellar hook assembly scaffolding protein (flgD) gene, complete cds	99	354

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

54	14	4872	4321	gb|U42012|	Treponema pallidum putative treponemal aqueous protein Tap1 (tap1) gene, and flagellar hook assembly scaffolding protein (flgD) gene, complete cds	98	189
54	15	4456	5550	gb|U42012|	Treponema pallidum putative treponemal aqueous protein Tap1 (tap1) gene, and flagellar hook assembly scaffolding protein (flgD) gene, complete cds	99	122
55	3	759	1289	gb|U97359|	Treponema pallidum 76K protein gene, complete cds	100	96
55	4	1256	1753	gb|U97359|	Treponema pallidum 76K protein gene, complete cds	99	494
55	5	2336	3370	gb|U97359|	Treponema pallidum 76K protein gene, complete cds	99	919
56	1	1	225	gb|U61851|	Treponema pallidum histidine kinase CheA (cheA), and chemotaxis proteins CheW (cheW), CheX (cheX) and CheY (cheY) genes, complete cds	100	138
56	2	129	308	gb|U61851|	Treponema pallidum histidine kinase CheA (cheA), and chemotaxis proteins CheW (cheW), CheX (cheX) and CheY (cheY) genes, complete cds	100	167
56	3	281	1669	gb|U61851|	Treponema pallidum histidine kinase CheA (cheA), and chemotaxis proteins CheW (cheW), CheX (cheX) and CheY (cheY) genes, complete cds	100	1389
56	4	1667	2155	gb|U61851|	Treponema pallidum histidine kinase CheA (cheA), and chemotaxis proteins CheW (cheW), CheX (cheX) and CheY (cheY) genes, complete cds	100	489
56	5	2128	2601	gb|U61851|	Treponema pallidum histidine kinase CheA (cheA), and chemotaxis proteins CheW (cheW), CheX (cheX) and CheY (cheY) genes, complete cds	98	397
63	13	5012	4779	gb|U12861|	Treponema pallidum Nichols TpN38(b) (tpn38(b)) gene, complete cds	100	186
63	14	5484	4945	gb|U12861|	Treponema pallidum Nichols TpN38(b) (tpn38(b)) gene, complete cds	100	458

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

63	15	5648	5409	gb|U12861|	Treponema pallidum Nichols TpN38(b) (tpn38(b)) gene, complete cds	100	190
63	22	11392	10127	gb|U02628|	Treponema pallidum pallidum Nichols TpN50 precursor (tpn50) gene, complete cds	99	1266
63	23	12291	11371	gb|L28427|	Treponema pallidum antigen (tpp57) gene, complete cds	99	312
64	9	3849	5198	gb|J04241|	T.pallidum 34 kd antigen gene, complete cds	99	177
64	10	5102	5317	gb|J04241|	T.pallidum 34 kd antigen gene, complete cds	100	216
64	11	5364	5795	gb|J04241|	T.pallidum 34 kd antigen gene, complete cds	96	382
64	12	6257	5658	gb|J04241|	T.pallidum 34 kd antigen gene, complete cds	93	414
64	13	5726	5965	gb|J04241|	T.pallidum 34 kd antigen gene, complete cds	100	200
64	14	6082	7029	gb|J04241|	T.pallidum 34 kd antigen gene, complete cds	96	404
68	1	512	3	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	97	482
69	6	2742	3845	gb|U61535|	Treponema pallidum gamma-glutamyl phosphate reductase (proA) and glutamate 5-kinase (proB) genes, complete cds	98	791
69	7	3578	4021	gb|U61535|	Treponema pallidum gamma-glutamyl phosphate reductase (proA) and glutamate 5-kinase (proB) genes, complete cds	99	428
69	8	3994	4596	gb|U61535|	Treponema pallidum gamma-glutamyl phosphate reductase (proA) and glutamate 5-kinase (proB) genes, complete cds	99	546
69	9	4491	4907	gb|U61535|	Treponema pallidum gamma-glutamyl phosphate reductase (proA) and glutamate 5-kinase (proB) genes, complete cds	100	370
69	10	4949	5296	gb|U61535|	Treponema pallidum gamma-glutamyl phosphate reductase (proA) and glutamate 5-kinase (proB) genes, complete cds	100	52
72	2	1313	312	gb|U97360|	Treponema pallidum HFLK homolog gene, complete cds	99	609
72	3	1137	2792	gb|U97360|	Treponema pallidum HFLK homolog gene,	99	553

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

					complete cds		
78	1	790	2	gb|U36839|	Treponema pallidum putative switch protein FliM (fliM) gene, partial cds, and FliY (fliY) gene, complete cds, export apparatus proteins FliP (fliP), FliQ (fliQ), FliR (fliR) and FlhB (flhB), signal transducing receptor FlhA (flhA), GTP binding protein >	99	716
78	2	1916	1074	gb|U36839|	Treponema pallidum putative switch protein FliM (fliM) gene, partial cds, and FliY (fliY) gene, complete cds, export apparatus proteins FliP (fliP), FliQ (fliQ), FliR (fliR) and FlhB (flhB), signal transducing receptor FlhA (flhA), GTP binding protein >	100	843
78	3	2320	1886	gb|U36839|	Treponema pallidum putative switch protein FliM (fliM) gene, partial cds, and FliY (fliY) gene, complete cds, export apparatus proteins FliP (fliP), FliQ (fliQ), FliR (fliR) and FlhB (flhB), signal transducing receptor FlhA (flhA), GTP binding protein >	100	373
78	4	2703	2224	gb|U36839|	Treponema pallidum putative switch protein FliM (fliM) gene, partial cds, and FliY (fliY) gene, complete cds, export apparatus proteins FliP (fliP), FliQ (fliQ), FliR (fliR) and FlhB (flhB), signal transducing receptor FlhA (flhA), GTP binding protein >	99	445
78	5	3709	2609	gb|U36839|	Treponema pallidum putative switch protein FliM (fliM) gene, partial cds, and FliY (fliY) gene, complete cds, export apparatus proteins FliP (fliP), FliQ (fliQ), FliR (fliR) and FlhB (flhB), signal transducing receptor FlhA (flhA), GTP binding protein >	99	721
78	6	3998	3645	gb|U28219|	Treponema pallidum flagellar hook (flgE), (orf4), flagellar motor protein (motA), flagellar motor	98	300

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

					protein (motB), (fliL), and flagellar switch protein (fliM) genes, complete cds, and (fliY) gene, partial cds		
78	7	4483	3707	gb|U28219|	Treponema pallidum flagellar hook (flgE), (orf4), flagellar motor protein (motA), flagellar motor protein (motB), (fliL), and flagellar switch protein (fliM) genes, complete cds, and (fliY) gene, partial cds	100	388
78	8	4500	4243	gb|U28219|	Treponema pallidum flagellar hook (flgE), (orf4), flagellar motor protein (motA), flagellar motor protein (motB), (fliL), and flagellar switch protein (fliM) genes, complete cds, and (fliY) gene, partial cds	100	126
81	6	3472	2957	gb|M17716|	T.pallidum basic membrane protein gene, complete cds	99	495
81	7	4116	3268	gb|M17716|	T.pallidum basic membrane protein gene, complete cds	100	612
81	8	6081	4114	gb|M17716|	T.pallidum basic membrane protein gene, complete cds	100	211
83	4	1058	1795	emb|X61227|TPTP70G	T.pallidum Tp70 gene (partial)	98	179
83	14	5091	5687	gb|U97362|	Treponema pallidum 22.5K protein gene, complete cds	100	80
83	15	5828	6454	gb|U97362|	Treponema pallidum 22.5K protein gene, complete cds	100	627
83	16	6617	7972	gb|U97362|	Treponema pallidum 22.5K protein gene, complete cds	97	77
84	1	3	152	gb|U61851|	Treponema pallidum histidine kinase CheA (cheA), and chemotaxis proteins CheW (cheW), CheX (cheX) and CheY (cheY) genes, complete cds	95	96
84	2	659	393	gb|U61851|	Treponema pallidum histidine kinase CheA (cheA), and chemotaxis proteins CheW (cheW), CheX (cheX) and CheY (cheY) genes, complete cds	100	106

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

84	3	837	598	gb|U70868|	Treponema pallidum ribosomal protein L28 homolog (rpl28) gene, complete cds	86	181
84	4	703	1236	gb|U70868|	Treponema pallidum ribosomal protein L28 homolog (rpl28) gene, complete cds	88	224
89	6	2962	4770	gb|U57757|	Treponema pallidum PolA gene, partial cds	99	330
89	7	4849	5748	gb|U57757|	Treponema pallidum PolA gene, partial cds	100	837
89	8	5678	8002	gb|U57757|	Treponema pallidum PolA gene, partial cds	99	2325
89	9	8378	8034	gb|U57757|	Treponema pallidum PolA gene, partial cds	99	201
99	1	87	1316	gb|L20301|	Treponema pallidum 38 kilodalton glucose/galactose binding lipoprotein (tpp38) gene, complete cds	99	1230
99	2	1379	2926	gb|U48416|	Treponema pallidum mgl operon: putative glucose/galactose binding protein (mglB), putative ATP binding protein (mglA), and hydrophobic putative membrane-associated protein (mglC) genes, complete cds	99	1548
99	3	2806	3483	gb|U45323|	Treponema pallidum putative ATP-binding protein (mglA) and putative methylgalactoside transport protein (mglC) genes, complete cds	99	624
99	4	3306	3043	gb|U45323|	Treponema pallidum putative ATP-binding protein (mglA) and putative methylgalactoside transport protein (mglC) genes, complete cds	98	264
99	5	3305	4534	gb|U48416|	Treponema pallidum mgl operon: putative glucose/galactose binding protein (mglB), putative ATP binding protein (mglA), and hydrophobic putative membrane-associated protein (mglC) genes, complete cds	99	1113
99	6	6086	4920	gb|U70661|	Treponema pallidum GTP-binding protein, mccF-like protein and ATP-dependent DNA helicase (RecG) genes, complete cds	99	1101
100	31	12458	13066	emb|X63965| TPFLAB13	T.pallidum endoflagellar genes flaB1 and flaB3	99	566
100	32	13027	13410	emb|X63965|	T.pallidum endoflagellar genes flaB1 and flaB3	100	370

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

				TPFLAB13			
100	33	13658	13957	emblX63965l TPFLAB13	T.pallidum endoflagellar genes f1aB1 and flaB3	100	262
100	34	13846	14544	emblX63965l TPFLAB13	T.pallidum endoflagellar genes f1aB1 and flaB3	98	628
104	7	2890	2552	gblM74825l	Treponema pallidum 17 kDa lipoprotein gene, complete cds	99	256
109	1	150	866	gblU88957l	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	100	143
112	5	3982	3023	gblU56999l	Treponema pallidum methyl-accepting chemotaxis protein (mcp-1) gene, complete cds, and potential regulatory molecule (pfoS/R) gene, partial cds	100	791
112	6	5935	4076	gblU56999l	Treponema pallidum methyl-accepting chemotaxis protein (mcp-1) gene, complete cds, and potential regulatory molecule (pfoS/R) gene, partial cds	100	1818
112	7	6555	5866	gblU56999l	Treponema pallidum methyl-accepting chemotaxis protein (mcp-1) gene, complete cds, and potential regulatory molecule (pfoS/R) gene, partial cds	99	477
112	8	7024	6545	gblU56999l	Treponema pallidum methyl-accepting chemotaxis protein (mcp-1) gene, complete cds, and potential regulatory molecule (pfoS/R) gene, partial cds	100	480
112	9	6572	7804	gblU56999l	Treponema pallidum methyl-accepting chemotaxis protein (mcp-1) gene, complete cds, and potential regulatory molecule (pfoS/R) gene, partial cds	100	939
112	10	7282	7022	gblU56999l	Treponema pallidum methyl-accepting chemotaxis protein (mcp-1) gene, complete cds, and potential regulatory molecule (pfoS/R) gene, partial cds	100	261
112	11	7819	7280	gblU56999l	Treponema pallidum methyl-accepting chemotaxis protein (mcp-1) gene, complete cds, and potential regulatory molecule (pfoS/R) gene, partial cds	100	231
112	17	13968	13516	emblX54111l TPTP4	T.pallidum Tp4 gene	99	449
120	1	850	44	gblU55214l	Treponema pallidum RuvB' gene, partial cds, and	100	779

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

					TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds		
121	23	12559	13242	gb|M10931|	T.pallidum tmpA gene encoding a membrane protein	98	51
121	24	13291	14418	gb|M10931|	T.pallidum tmpA gene encoding a membrane protein	99	1128
121	25	14361	14804	gb|M58562|	Treponema pallidum tmpB gene, complete cds	100	375
121	26	14750	15178	gb|M58562|	Treponema pallidum tmpB gene, complete cds	99	342
132	1	2	724	gb|M88769|	Treponema pallidum 47-kilodalton antigen gene, complete cds	99	723
133	1	709	2	gb|U73748|	Treponema pallidum putative aspartate aminotransferase TpAAT (tpaat) and leucine-rich repeat protein TpLRR genes, complete cds	99	351
133	2	1346	888	gb|U73748|	Treponema pallidum putative aspartate aminotransferase TpAAT (tpaat) and leucine-rich repeat protein TpLRR genes, complete cds	98	328
133	3	2265	1549	gb|U73748|	Treponema pallidum putative aspartate aminotransferase TpAAT (tpaat) and leucine-rich repeat protein TpLRR genes, complete cds	98	549
142	3	4022	4951	gb|U78094|	Treponema pallidum DNA gyrase subunit B (gyrB) gene, complete cds	100	92
142	4	5107	6972	gb|U78094|	Treponema pallidum DNA gyrase subunit B (gyrB) gene, complete cds	99	1855
145	12	3527	3958	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	96	98
153	2	1801	746	gb|U32683|	Treponema pallidum cytoplasmic filament protein A (cfpA) gene, complete cds	100	423
166	1	19	843	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds	99	825
166	2	689	994	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and	98	284

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

					Tpp15 genes, complete cds		
166	3	943	1710	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds	99	750
166	4	1708	2655	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds	98	915
166	5	2594	2821	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds	98	164
166	6	2814	3287	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds	100	474
166	7	3352	3831	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds	100	456
166	8	3714	4061	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds	99	254
168	1	1113	4	gb|U36839|	Treponema pallidum putative switch protein FliM (fliM) gene, partial cds, and FliY (fliY) gene, complete cds, export apparatus proteins FliP (fliP), FliQ (fliQ), FliR (fliR) and FlhB (flhB), signal transducing receptor FlhA (flhA), GTP binding protein >	99	1043
172	3	1442	498	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	97	98
172	4	519	989	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	84	119
174	2	1640	372	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	90	845
176	1	71	835	gb|U28219|	Treponema pallidum flagellar hook (flgE), (orf4), flagellar motor protein (motA), flagellar motor	99	681

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

					protein (motB), (fliL), and flagellar switch protein (fliM) genes, complete cds, and (fliY) gene, partial cds		
176	2	899	1228	gb|U28219|	Treponema pallidum flagellar hook (flgE), (orf4), flagellar motor protein (motA), flagellar motor protein (motB), (fliL), and flagellar switch protein (fliM) genes, complete cds, and (fliY) gene, partial cds	100	219
190	1	3	383	gb|U61851|	Treponema pallidum histidine kinase CheA (cheA), and chemotaxis proteins CheW (cheW), CheX (cheX) and CheY (cheY) genes, complete cds	98	211
190	2	248	754	gb|U61851|	Treponema pallidum histidine kinase CheA (cheA), and chemotaxis proteins CheW (cheW), CheX (cheX) and CheY (cheY) genes, complete cds	99	441
205	1	2	568	gb|U65743|	Treponema pallidum outer membrane protein (tromp2) gene, complete cds	98	565
206	2	528	1214	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds	98	177
206	3	1183	740	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds	98	146
206	4	1025	804	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds	98	63
206	5	1599	1204	gb|U55214|	Treponema pallidum RuvB' gene, partial cds, and TroA, TroB, TroC, TroD, TroR, Pgm, Pfs, and Tpp15 genes, complete cds	99	396
223	1	1	564	emb|X54111| TPTP4	T.pallidum Tp4 gene	99	558
224	3	992	453	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	97	411
224	4	486	731	gb|U88957|	Treponema pallidum major outer sheath protein	100	154

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

					homolog Msp (msp) gene, complete cds		
231	1	447	46	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	100	96
491	1	394	83	gb|U95214|	Treponema pallidum octaprenyl-diphosphate synthase gene, complete cds	97	255
495	1	406	2	gb|U61851|	Treponema pallidum histidine kinase CheA (cheA), and chemotaxis proteins CheW (cheW), CheX (cheX) and CheY (cheY) genes, complete cds	100	368
495	2	582	235	gb|U61851|	Treponema pallidum histidine kinase CheA (cheA), and chemotaxis proteins CheW (cheW), CheX (cheX) and CheY (cheY) genes, complete cds	99	201
562	1	718	44	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	99	538
587	1	110	439	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	95	201
592	1	3	410	gb|J04241|	T.pallidum 34 kd antigen gene, complete cds	93	315
602	1	1	246	gb|U56999|	Treponema pallidum methyl-accepting chemotaxis protein (mcp-1) gene, complete cds, and potential regulatory molecule (pfoS/R) gene, partial cds	98	177
626	2	642	73	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	72	325
627	1	402	88	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	100	231
627	2	911	366	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	89	495
630	1	1	549	gb|U56999|	Treponema pallidum methyl-accepting chemotaxis protein (mcp-1) gene, complete cds, and potential regulatory molecule (pfoS/R) gene, partial cds	100	393
630	2	165	4	gb|U56999|	Treponema pallidum methyl-accepting chemotaxis protein (mcp-1) gene, complete cds, and potential regulatory molecule (pfoS/R) gene, partial cds	100	162
633	2	102	332	gb|U88957|	Treponema pallidum major outer sheath protein	100	129

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

					homolog Msp (msp) gene, complete cds		
640	1	708	310	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	93	274
645	1	3	311	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	85	122
646	3	896	501	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	91	148
653	1	1	282	emb|X63965| TPFLAB13	T.pallidum endoflagellar genes f1aB1 and flaB3	97	126
653	2	499	161	emb|X63965| TPFLAB13	T.pallidum endoflagellar genes f1aB1 and flaB3	98	272
654	1	1	552	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	68	293
654	2	624	4	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	68	293
660	1	585	4	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	71	252
665	1	434	3	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	84	119
665	2	57	434	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	84	119
666	1	2	373	gb|U97358|	Treponema pallidum 29K protein gene, complete cds	98	178
666	2	496	119	gb|U97358|	Treponema pallidum 29K protein gene, complete cds	98	301
669	2	412	71	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	75	203
671	1	361	62	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	100	297
672	1	3	653	gb|U88957|	Treponema pallidum major outer sheath protein homolog Msp (msp) gene, complete cds	90	419
681	1	109	258	gb|U78094|	Treponema pallidum DNA gyrase subunit B (gyrB)	94	118

TABLE 1.

Treponema pallidum - Coding regions containing know sequences

					gene, complete cds		
696	1	304	483	gb|U45323|	Treponema pallidum putative ATP-binding protein (mglA) and putative methylgalactoside transport protein (mglC) genes, complete cds	97	129

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

Contig ID	ORF ID	Start (nt)	Stop (nt)	match acession	match gene name	% sim	% ident
40	1	540	316	gp|X57972|	beta lactamase TEM6 [Escherichia coli] >pir|S|S24415 beta-lactamase (EC 3.5.2.6) TEM6 - Escherichia coli	100	100
40	2	965	405	gp|M74750|	beta-lactamase [Escherichia coli] >gp|D14640|ECORPL12A_2 beta-lactamase [Plasmid pKF2] >gp|U36911|SAU36911_1 beta-lactamase [Staphylococcus aureus] >gp|U36912|SAU36912_1 beta-lactamase [Staphylococcus aureus] >gp|U25060|XXU25060_3 beta lactamase [unidenti	100	99
61	1	418	972	gp|M74750|	beta-lactamase [Escherichia coli] >gp|D14640|ECORPL12A_2 beta-lactamase [Plasmid pKF2] >gp|U36911|SAU36911_1 beta-lactamase [Staphylococcus aureus] >gp|U36912|SAU36912_1 beta-lactamase [Staphylococcus aureus] >gp|U25060|XXU25060_3 beta lactamase [unidenti	100	98
96	1	12	275	pir|S|A43562	homeotic protein Hox 4.3 - mouse	100	85
228	2	832	1425	gp|M74750|	beta-lactamase [Escherichia coli] >gp|D14640|ECORPL12A_2 beta-lactamase [Plasmid pKF2] >gp|U36911|SAU36911_1 beta-lactamase [Staphylococcus aureus] >gp|U36912|SAU36912_1 beta-lactamase [Staphylococcus aureus] >gp|U25060|XXU25060_3 beta lactamase [unidenti	100	100
534	1	107	616	gp|J02459|	Bacteriophage lambda, complete genome. [Bacteriophage lambda] >pir|S|VHBPEL major capsid protein E - phage lambda	100	100
579	1	2	241	gp|X66453|	DNA-directed RNA polymerase [Euplotes octocarinatus]	100	87
637	1	29	394	gp|A04190|	galactosidase fusion protein [unidentified] >gp|K01075|SYNCSPBL_1 Plasmodium knowlesi circumsporozoite protein repeat region fused with beta-lactamase gene of pBR322. [Artificial gene] {SUB 181-300} >gp|L30112|P14BLAREP_1 beta-lactamase [Plasmid pR01614]	100	99
45	16	15198	14485	pir|S|A37053	33K endoflagellar protein FlaB2 - Treponema pallidum >pir|S|D32351 33K class B flagellar protein, periplasmic - Treponema pallidum subsp. pallidum (fragment) {SUB 1-21}	98	98
607	1	223	2	gp|U13869|	lacZ alpha peptide [Cloning vector]	98	98

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

642	1	81	392	gp|M74750|	beta-lactamase [Escherichia coli] >gp|D14640|ECORPL12A_2 beta-lactamase [Plasmid pKF2] >gp|U36911|SAU36911_1 beta-lactamase [Staphylococcus aureus] >gp|U36912|SAU36912_1 beta-lactamase [Staphylococcus aureus] >gp|U25060|XXU25060_3 beta lactamase [unidenti	98	98
14	1	3	329	gp|U25059|	LacZ alpha peptide [unidentified] >gp|U25060|XXU25060_2 LacZ alpha peptide [unidentified] >gp|U25061|XXU25061_2 LacZ alpha peptide [unidentified] >gp|U23751|CVU23751_2 beta galactosidase [Cloning vector pBBR1MCS-2]	96	96
101	12	5676	5107	gp|Z67753|	DNA-replication helicase [Chloroplast Odontella sinensis]	96	70
169	1	2	190	gp|V00083|	Artificial cloning vehicle pBR327, derived from pBR322. The sequence was not resequenced but deduced from the pBR322 sequence. Contains the reading frames for ampicillin resistance (Apr) and tetracycline resistance (Tcr) and an origin of replication. [uni	96	96
588	1	1	213	gp|X94607|	Mef1 protein [Saccharomyces cerevisiae]	96	74
11	15	6934	8265	gp|X04581|	E.coli recB gene for exonuclease V. [Escherichia coli] >gp|U29581|ECU29581_31 exonuclease V subunit [Escherichia coli] >gp|U29581|ECU29581_31 exonuclease V subunit [Escherichia coli] >pir|S|NCECX5 exodeoxyribonuclease V (EC 3.1.11.5) 135K chain - Escheric	95	65
34	1	328	2	gp|U37456|	beta-lactamase [Cloning vector YITAG100] >gp|U37457|CVU37457_2 beta-lactamase [Cloning vector YATAG200]	95	94
673	1	376	2	gp|A04190|	galactosidase fusion protein [unidentified] >gp|K01075|SYNCSPBL_1 Plasmodium knowlesi circumsporozoite protein repeat region fused with beta-lactamase gene of pBR322. [Artificial gene] {SUB 181-300} >gp|L30112|P14BLAREP_1 beta-lactamase [Plasmid pR01614]	95	95
79	9	4159	5235	gp|U15609|	flagellar switch protein [Treponema denticola] >gp|L36851|TRPFLIG_1 fliG gene product [Treponema denticola]	94	85
117	2	1565	921	gp|D64006|	hypothetical protein [Synechocystis sp.]	93	76
16	25	15399	16082	gp|L10328|	gidB (CG Site No. 18274) gene product [Escherichia coli]	92	71

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					>gp|X01631|ECUNC_3 E. coli origin of replication oriC and genes gid, unc, EcoURF-1 and glmS. [Escherichia coli] >pir|S|BVECQB gidB protein - Escherichia coli		
42	27	18334	18720	gp|M57776|	ribosomal protein S12 [Leptospira biflexa] >pir|S|A36152 ribosomal protein S12 - Leptospira biflexa (serotype patoc)	91	82
197	1	1	393	gp|D64000|	hypothetical protein [Synechocystis sp.]	91	81
547	1	552	370	gp|U19581|	unknown [Klebsiella pneumoniae]	91	81
16	49	29062	28667	gp|M24466|	S.typhimurium flagellar L-ring (flgH), flagellar P-ring (flgI), and flagellar (flgJ) genes, complete cds. [Salmonella typhimurium] >pir|S|C30930 flagellar basal body protein flgJ - Salmonella typhimurium	90	55
650	1	179	391	gp|U17139|	Phagemid cloning vector pSIT, complete sequence. [Cloning vector pSIT] >gp|U47102|CVU47102_2 beta-lactamase [Cloning vector pALTER<R>-Ex1]	90	89
4	8	6487	5276	gp|X52898|	glyceraldehyde 3-phosphate dehydrogenase [Trypanosoma cruzi] >gp|X52898|TCGAP_2 glyceraldehyde 3-phosphate dehydrogenase [Trypanosoma cruzi] >pir|S|DEUT1C glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12), glycosomal - Trypanosoma cruzi	89	72
8	45	26587	27414	gp|D30690|	HSP40 [Staphylococcus aureus]	89	68
25	1	224	3	pir|S|A60163	glycoprotein IIb - rat	88	77
64	8	3423	3635	gp|Z19059|	Cek8 protein [Gallus gallus] >gp|Z19059|GGCEK8A_1 Cek8 protein [Gallus gallus] >pir|S|S33505 protein-tyrosine kinase (EC 2.7.1.112) Cek8 - chicken (fragment) {SUB 2-849} >gp|X57241|MMMPK3_1 tyrosine kinase [Mus musculus] {SUB 611-668} >pir|S|PT0183 protei	88	66
598	1	361	212	gp|L45070|	glutamine amidotransferase [Haemophilus influenzae] >gp|U32726|HIU32726_10 glutamine amidotransferase [Haemophilus influenzae] >gp|U00073|HIU00073_51 glutamine amidotransferase [Haemophilus influenzae] >gp|U32835|HIU32835_5 glutamine amidotransferase [Hae	88	72

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

8	17	8697	9533	gp|L36380|	tufA gene product [Neisseria gonorrhoeae]	87	76
16	27	16417	16127	gp|F14628|	cyclophilin B [Sus scrofa]	87	79
30	9	4454	4765	gp|U13165|	sigma 43 subunit of RNA polymerase [Listeria monocytogenes]	87	50
44	16	9669	9334	gp|L45205|	asparagine synthetase A [Haemophilus influenzae] >gp|U32738|HIU32738_3 asparagine synthetase A [Haemophilus influenzae] >gp|U00074|HIU00074_84 asparagine synthetase A [Haemophilus influenzae] >gp|U32847|HIU32847_3 asparagine synthetase [Haemophilus influe	87	76
143	2	950	768	gp|S65735|	membrane glycoprotein M6 [Mus sp.]	87	62
42	22	15120	15563	gp|M38305|	E.coli RNA polymerase beta subunit (rpoC) gene, partial cds. [Escherichia coli]	86	72
48	18	11952	11773	gp|L45145|	periplasmic ribose-binding protein [Haemophilus influenzae] >gp|U32732|HIU32732_5 periplasmic ribose-binding protein [Haemophilus influenzae] >gp|U00074|HIU00074_25 periplasmic ribose-binding protein [Haemophilus influenzae] >gp|U32841|HIU32841_5 D-ribose	86	50
113	2	1687	2058	gp|Z49124|	SecA [Chloroplast Spinacia oleracea]	86	65
113	4	2604	2900	gp|X55034|	SecA protein [Escherichia coli] >gp|D10483|ECO110K_78 secA protein [Escherichia coli] >gp|M20791|ECOSECA_2 secA gene product [Escherichia coli] >pir|S|A31088 secA protein - Escherichia coli >gp|M19211|ECOENVAA_4 secA gene product [Escherichia coli] {SUB 1	86	65
117	3	2176	1472	gp|L45353|	ATP-dependent clp protease proteolytic component [Haemophilus influenzae] >gp|U32754|HIU32754_5 ATP-dependent clp protease proteolytic component [Haemophilus influenzae] >gp|U00076|HIU00076_53 ATP-dependent clp protease proteolytic component [Haemophilus	86	65
8	19	9481	10374	gp|J01690|	elongation factor Tu [Escherichia coli] >gp|U18997|ECOUW67_263 tufA gene product [Escherichia coli] >pir|S|EFECTA translation elongation factor Tu.A - Escherichia coli	85	68
63	12	4803	4378	gp|U14003|	Escherichia coli K-12 chromosomal region from 92.8 to 00.1	85	74

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					minutes. [Escherichia coli] >pir|S|S56315 hypothetical protein 510 - Escherichia coli >gp|U00006|ECOUW89_134 E. coli chromosomal region from 89.2 to 92.8 minutes. [Escherichia coli] {SUB 9-510}		
588	2	557	117	gp|L14925|	nitrogen fixation protein [Anabaena sp.]	85	72
693	1	64	441	gp|M96436|	oxaloacetate decarboxylase [Klebsiella pneumoniae] >pir|S|A44464 oxaloacetate decarboxylase beta subunit (C terminus) - Klebsiella pneumoniae (fragment) >gp|M26290|KPNOADBG2_1 oadB gene product [Klebsiella pneumoniae] {SUB 4-141}	85	82
1	5	1593	2015	gp|L44957|	Holliday junction DNA helicase [Haemophilus influenzae] >gp|U32716|HIU32716_13 Holliday junction DNA helicase [Haemophilus influenzae] >gp|U00072|HIU00072_42 Holliday junction DNA helicase [Haemophilus influenzae] >gp|U32825|HIU32825_12 Holliday junction	84	63
21	20	11250	11630	gp|Z46729|	unknown [Saccharomyces cerevisiae] >pir|S|S49802 hypothetical protein YM9958.03c - yeast (Saccharomyces cerevisiae)	84	61
47	25	14251	13676	pir|S|C47154	ribosomal protein S16 - Bacillus subtilis	84	59
49	1	1	1092	gp|X89411|	SecA [Rhodobacter capsulatus]	84	70
79	11	6479	7015	gp|M72718|	B.subtilis flaA locus operon. [Bacillus subtilis] >gp|X56049|BSFLAAO_3 B. subtilis FlaA locus operon. [Bacillus subtilis] >pir|S|PWBSAS H+-transporting ATP synthase alpha chain homolog - Bacillus subtilis	84	68
87	2	1995	1639	gp|Z54171|	elongation factor EF-G [Rickettsia prowazekii] >gp|U02603|RPU02603_8 elongation factor EF-G [Rickettsia prowazekii] {SUB 1-47}	84	69
101	23	15080	13362	gp|U43536|	clpB gene product [Corynebacterium glutamicum]	84	68
8	43	24498	26312	gp|X67646|	heat-shock protein [Borrelia burgdorferi] >gp|M96847|BORGRPEPLS_2 dnaK homologue gene product [Borrelia burgdorferi] >gp|M97912|BORHSP70A_1 70 kDa heat shock protein [Borrelia burgdorferi] >gp|S42385|S42385_1 HSP70 homolog [Borrelia burgdorferi, CA12 isol	83	70

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

10	21	12836	11373	gp|U35673|	Rho [Borrelia burgdorferi] >pir|S|S35618 rho protein - Lyme disease spirochete (fragment) {SUB 97-515} >gp|L07656|BORRHO_1 Rho protein [Borrelia burgdorferi] {SUB 127-515}	83	64
23	6	2510	3373	gp|M64730|	DNA mismatch repair protein [Escherichia coli] >gp|U29579|ECU29579_30 DNA mismatch repair protein [Escherichia coli]	83	66
45	1	951	4	gp|M77351|	ATP-binding protein [Streptococcus mutans] >pir|S|E42400 ATP-binding protein MsmK - Streptococcus mutans >pir|S|C27626 hypothetical protein 2 - Streptococcus mutans (fragment) {SUB 1-33}	83	67
61	14	8382	7999	gp|U45426|	heat shock protein HTPG [Borrelia burgdorferi] >gp|L32145|BORHTPG_1 C62.5 heat shock protein [Borrelia burgdorferi] {SUB 497-575}	83	56
81	5	2415	2224	gp|X90857|	-14 gene product [Homo sapiens]	83	50
115	6	2113	2598	gp|X91655|	lepA gene product [Bacillus subtilis] >gp|D17650|BACGPR_4 ORF80 protein [Bacillus subtilis] {SUB 1-327}	83	61
41	9	2379	3092	gp|D10279|	ORF3 [Bacillus subtilis] >pir|S|JN0146 hypothetical protein (div+ 3' region) - Bacillus subtilis (fragment)	82	67
81	1	287	628	gp|L10132|	groES gene product [Bacillus stearothermophilus] >pir|S|A49855 heat shock protein GroES - Bacillus stearothermophilus	82	60
165	6	1887	2087	gp|U32690|	permease [Haemophilus influenzae]	82	54
8	20	10445	10771	gp|L45414|	ribosomal protein S10 [Haemophilus influenzae] >gp|U32761|HIU32761_6 ribosomal protein S10 [Haemophilus influenzae] >gp|U00077|HIU00077_35 ribosomal protein S10 [Haemophilus influenzae] >gp|U32707|HIU32707_9 30S ribosomal protein S10 [Haemophilus influenz	81	57
8	41	24119	23823	gp|S77354|	VP1 [Unknown.]	81	54
44	17	10376	9630	gp|L45205|	asparagine synthetase A [Haemophilus influenzae] >gp|U32738|HIU32738_3 asparagine synthetase A [Haemophilus influenzae] >gp|U00074|HIU00074_84 asparagine synthetase A [Haemophilus influenzae] >gp|U32847|HIU32847_3 asparagine	81	70

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					synthetase [Haemophilus influe		
47	1	900	4	gp|D49951|	H2O-forming NADH Oxidase [Streptococcus mutans]	81	64
54	28	8744	8499	gp|U39684|	purine-nucleoside phosphorylase [Mycoplasma genitalium] >pir|S|D64205 purine-nucleoside phosphorylase - Mycoplasma genitalium (SGC3)	81	53
57	11	7418	9556	gp|L77216|	Lon protease [Borrelia burgdorferi]	81	59
57	12	9477	9857	gp|J03896|	E.coli ATP-dependent protease La (lon) gene, complete cds. [Escherichia coli]	81	66
101	25	16036	14795	gp|U43536|	clpB gene product [Corynebacterium glutamicum]	81	69
112	1	912	4	gp|Z12160|	gidA gene product [Borrelia burgdorferi] >gp|Z12160|BBGIDAG_1 division protein [Borrelia burgdorferi] {SUB 529-593} >gp|X95669|BBTHDFGID_2 gidA gene product [Borrelia burgdorferi] {SUB 1-29} >gp|X95668|BBGIDMOXR_1 gidA gene product [Borrelia burgdorferi]	81	64
174	1	117	317	gp|X15540|	T.brucei pgi gene for glucose-6-phosphate isomerase (EC 5.3.1.9). [Trypanosoma (Trypanozoon) brucei] >pir|S|NUUTB glucose-6-phosphate isomerase (EC 5.3.1.9) - Trypanosoma brucei	81	63
337	1	84	374	gp|M25927|	Influenza A/swine/Hong Kong/126/82 (H3N2) PB1 gene, complete cds. [Influenza virus type A]	81	54
555	1	2	190	gp|L14925|	nitrogen fixation protein [Anabaena sp.]	81	79
683	1	388	170	gp|X69618|	inhibin alpha-subunit [Mus musculus] >pir|S|S31439 inhibin alpha chain - mouse	81	72
8	28	16269	16844	pir|S|R5BS5F	ribosomal protein L5 - Bacillus stearothermophilus	80	57
16	16	11479	10670	gp|X77515|	pyruvate oxidoreductase [Rhodospirillum rubrum] >pir|S|S41961 pyruvate oxidoreductase - Rhodospirillum rubrum	80	67
17	9	4729	4893	gp|L11706|	hormone-sensitive lipase [Homo sapiens]	80	80
35	16	9193	9642	gp|L04500|	thioredoxin reductase [Eubacterium acidaminophilum] >pir|S|S38988 thioredoxin reductase chain B - Eubacterium acidaminophilum	80	59
36	16	10409	10756	gp|X78993|	probable transfer RNA-Gly synthetase [Saccharomyces cerevisiae]	80	55

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					>gp|Z35990|SCYBR121C_1 GSR1 gene product [Saccharomyces cerevisiae] >pir|S|S48285 probable glycine--tRNA ligase (EC 6.1.1.14) GRS1 - yeast (Saccharomyces cerevisiae)		
42	19	11240	13345	gp|L48488|	RNA polymerase beta subunit [Borrelia burgdorferi]	80	67
42	20	13225	13983	gp|L48488|	RNA polymerase beta subunit [Borrelia burgdorferi]	80	72
42	21	13981	15210	gp|L43593|	RNA polymerase beta' subunit [Bacillus subtilis] >gp|L43593|BACBPSO_1 RNA polymerase beta' subunit [Bacillus subtilis] >gp|L24376|BACRPLL_4 RNA polymerase beta'-subunit [Bacillus subtilis] {SUB 1-466}	80	65
45	3	2471	5083	gp|M31045|	E.coli ATP-dependent C1p protease (clpA) gene, complete cds. [Escherichia coli] >pir|S|SUECCA ATP-dependent Clp proteinase (EC 3.4.21.-) chain A - Escherichia coli >gp|M23220|ECOCLPA_1 ATP-dependent protease [Escherichia coli] {SUB 1-28}	80	62
49	17	10127	9942	gp|Z22606|	H(+)-transporting ATP synthase [Streptomyces lividans] >pir|S|S37545 H+-transporting ATP synthase (EC 3.6.1.34) alpha chain - Streptomyces lividans	80	50
54	19	6858	6424	gp|Z35953|	MIS1 gene product [Saccharomyces cerevisiae] >gp|J03724|YSCMIS1A_1 S.cerevisiae mitochondrial C-1-Tetrahydrofolate synthase gene (MIS1). [Saccharomyces cerevisiae] >pir|S|A28174 C1-tetrahydrofolate synthase precursor, mitochondrial - yeast (Saccharomyces	80	52
101	7	2602	3246	gp|M96434|	oxaloacetate decarboxylase [Salmonella typhimurium] >pir|S|B44465 sodium ion pump oxaloacetate decarboxylase subunit alpha - Salmonella typhimurium	80	64
116	10	4191	3247	gp|U32847|	ribosomal protein S1 homolog, RNA-binding protein [Haemophilus influenzae]	80	63
170	3	2283	1651	gp|U43739|	FtsZ [Borrelia burgdorferi]	80	67
561	1	112	555	gp|L45199|	glucose-6-phosphate 1-dehydrogenase [Haemophilus influenzae] >gp|U32737|HIU32737_8 glucose-6-phosphate 1-dehydrogenase [Haemophilus influenzae] >gp|U00074|HIU00074_79 glucose-6-phosphate 1-dehydrogenase [Haemophilus influenzae] >gp|U32846|HIU32846_7 gluco	80	69

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

573	1	234	452	gp|U00019|	B2235_C2_195 [Mycobacterium leprae]	80	65
5	10	6650	7417	gp|D26562|	'plasmid copy number control protein pcnB' [Escherichia coli] >pir|S|S45212 plasmid copy number control protein - Escherichia coli	79	58
27	5	2427	2768	gp|Z49782|	peptide chain release factor 1 [Bacillus subtilis] >pir|S|S55437 peptide chain release factor 1 - Bacillus subtilis	79	61
42	17	9764	10186	gp|X53072|	S. typhimurium rplJ and rplL genes for ribosomal protein L10 and L7/L12. [Salmonella typhimurium] >pir|S|R5EB12 ribosomal protein L7/L12 - Salmonella typhimurium	79	61
57	2	428	1900	gp|J04836|	M.barkeri ATPase alpha and beta subunit (atpA and atpB) genes, complete cds. [Methanosarcina barkeri] >pir|S|A34283 H+-transporting ATP synthase (EC 3.6.1.34) alpha chain - Methanosarcina barkeri	79	63
60	6	4417	3140	gp|J04836|	M.barkeri ATPase alpha and beta subunit (atpA and atpB) genes, complete cds. [Methanosarcina barkeri] >pir|S|A34283 H+-transporting ATP synthase (EC 3.6.1.34) alpha chain - Methanosarcina barkeri	79	64
85	4	1001	1450	gp|L45451|	carbon storage regulator [Haemophilus influenzae] >gp|U32763|HIU32763_8 carbon storage regulator [Haemophilus influenzae] >gp|U00077|HIU00077_72 carbon storage regulator [Haemophilus influenzae] >gp|U32709|HIU32709_9 carbon storage regulator [Haemophilus	79	49
123	2	1158	550	gp|M80215|	uvs402 protein [Streptococcus pneumoniae] >pir|S|A42385 uvr-402 protein - Streptococcus pneumoniae plasmid pSB470	79	62
165	5	1287	1766	gp|L45262|	ATP-binding protein [Haemophilus influenzae] >gp|U32744|HIU32744_7 ATP-binding protein [Haemophilus influenzae] >gp|U00075|HIU00075_40 ATP-binding protein [Haemophilus influenzae] >gp|U32690|HIU32690_8 transport ATPase [Haemophilus influenzae] >pir|S|C640	79	54
6	2	1617	430	gp|X15867|	Micrococcus luteus homolog of the E.coli uvrA gene. [Micrococcus luteus] >pir|S|S04781 uvrA protein - Micrococcus luteus	78	65
21	11	7812	9110	gp|L08854|	valyl-tRNA synthetase [Lactobacillus casei] >pir|S|A49856 valine--	78	58

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					tRNA ligase (EC 6.1.1.9) - Lactobacillus casei		
22	21	14655	14203	gp|U00039|	E. coli chromosomal region from 76.0 to 81.5 minutes. [Escherichia coli] >pir|S|S47728 yhiD protein - Escherichia coli >gp|D11109|ECO10KLS_4 ORF-C [Escherichia coli] {SUB 1-148}	78	66
29	24	12039	10612	gp|D64006|	hypothetical protein [Synechocystis sp.]	78	62
44	8	4761	4459	gp|L25421|	phosphatase [Treponema denticola]	78	64
84	9	5019	4426	gp|U20445|	BirA protein [Bacillus subtilis]	78	47
100	14	3613	3365	gp|L45070|	glutamine amidotransferase [Haemophilus influenzae] >gp|U32726|HIU32726_10 glutamine amidotransferase [Haemophilus influenzae] >gp|U00073|HIU00073_51 glutamine amidotransferase [Haemophilus influenzae] >gp|U32835|HIU32835_5 glutamine amidotransferase [Hae	78	61
122	7	3432	4901	gp|M88581|	transfer RNA-Leu synthetase [Bacillus subtilis] >pir|S|A41882 leucine--tRNA ligase (EC 6.1.1.4) - Bacillus subtilis	78	67
172	1	26	277	gp|X53456|	plasma membrane Ca2+ pump (PMCA1b) [Sus scrofa] >pir|S|S13057 Ca2+-transporting ATPase (EC 3.6.1.38) - pig	78	42
726	1	421	140	gp|U34201|	cardiac triadin isoform 3 [Oryctolagus cuniculus]	78	69
10	19	10346	9219	gp|M37487|	protein D [Haemophilus influenzae] >gp|A28485|A28485_1 protein D [Haemophilus influenzae] >pir|S|A43576 protein D precursor - Haemophilus influenzae	77	58
30	13	8251	9300	gp|M96343|	MreB protein [Bacillus subtilis]	77	53
36	20	13764	16661	gp|U30472|	DNA polymerase III holoenzyme alpha subunit [Vibrio cholerae]	77	62
42	26	17880	18242	gp|M22622|	L.biflexa acetylornithine deacetylase (argE; complete cds), and ribosomal subunit protein (rpsL; 5' end) genes. [Leptospira biflexa] >pir|S|A31840 RNA polymerase beta chain homolog - Leptospira biflexa (serotype patoc)	77	60
45	4	3305	3072	gp|U15146|	polyprotein 1b [Lactate dehydrogenase-elevating virus]	77	66
70	1	3	890	gp|U10927|	CapD [Staphylococcus aureus]	77	60
74	15	8192	9343	gp|U36839|	orf304 gene product [Treponema pallidum]	77	47
101	10	3523	4605	pir|S|B36505	oxaloacetate decarboxylase (EC 4.1.1.3) beta chain - Klebsiella pneumoniae	77	60

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

115	8	3077	3955	gp|X91655|	lepA gene product [Bacillus subtilis] >gp|D17650|BACGPR_4 ORF80 protein [Bacillus subtilis] {SUB 1-327}	77	54
121	19	10031	8298	gp|U01376|	ATP-binding protein [Escherichia coli] >gp|U01376|U01376_4 ATP-binding protein [Escherichia coli] >gp|M83138|ECOFTSHJA_2 ftsH gene product [Escherichia coli] {SUB 4-647}	77	58
137	3	1603	92	gp|M22039|	CTP synthetase [Bacillus subtilis] >gp|Z49782|BSDNA320D_10 CTP synthase [Bacillus subtilis] >pir|S|SYBSTP CTP synthase (EC 6.3.4.2) - Bacillus subtilis	77	54
160	1	3	338	gp|Z12160|	gidA gene product [Borrelia burgdorferi] >gp|Z12160|BBGIDAG_1 division protein [Borrelia burgdorferi] {SUB 529-593} >gp|X95669|BBTHDFGID_2 gidA gene product [Borrelia burgdorferi] {SUB 1-29} >gp|X95668|BBGIDMOXR_1 gidA gene product [Borrelia burgdorferi]	77	60
195	1	388	2	gp|M64730|	DNA mismatch repair protein [Escherichia coli] >gp|U29579|ECU29579_30 DNA mismatch repair protein [Escherichia coli]	77	59
8	10	5202	5498	gp|U15186|	sua5 [Mycobacterium leprae]	76	58
12	3	1872	2051	gp|M17282|	elastin [Homo sapiens]	76	61
12	10	5367	5642	gp|L45661|	transketolase 1 (TK 1) [Haemophilus influenzae] >gp|U32783|HIU32783_2 transketolase 1 (TK 1) [Haemophilus influenzae] >gp|U00079|HIU00079_82 transketolase 1 (TK 1) [Haemophilus influenzae] >gp|U32729|HIU32729_2 transketolase 2 [Haemophilus influenzae] >pi	76	60
19	18	13976	13413	gp|L77246|	yppQ gene product [Bacillus subtilis]	76	53
28	3	1978	1724	gp|X78206|	glucose-6-phosphate isomerase [Leishmania mexicana]	76	52
42	18	10347	11399	gp|L48488|	RNA polymerase beta subunit [Borrelia burgdorferi]	76	62
42	28	18795	19211	gp|L45221|	ribosomal protein S7 [Haemophilus influenzae] >gp|U32739|HIU32739_9 ribosomal protein S7 [Haemophilus influenzae] >gp|U00074|HIU00074_100 ribosomal protein S7 [Haemophilus influenzae] >gp|U32848|HIU32848_9 30S ribosomal protein S7 [Haemophilus influenzae]	76	61

PCT/98/13041

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

45	14	11488	12753	gp|U28377|	metK gene product [Escherichia coli]	76	64
57	9	6841	7014	gp|L00673|	tyrosine aminotransferase [Trypanosoma cruzi]	76	53
83	3	870	1112	gp|X60343|	glyceraldehyde 3-phosphate dehydrogenase [Hordeum vulgare] >pir|S|DEBHG glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12) - barley	76	35
89	3	1288	1584	gp|M30198|	recQ gene product [Escherichia coli]	76	55
100	11	2971	2693	gp|U37523|	glucosamine synthetase [Sphingomonas yanoikuyae] >gp|U37523|SYU37523_3 glucosamine synthetase [Sphingomonas yanoikuyae]	76	65
101	11	4526	4879	gp|M96436|	oxaloacetate decarboxylase [Klebsiella pneumoniae] >pir|S|A44464 oxaloacetate decarboxylase beta subunit (C terminus) - Klebsiella pneumoniae (fragment) >gp|M26290|KPNOADBG2_1 oadB gene product [Klebsiella pneumoniae] {SUB 4-141}	76	61
150	16	12992	13774	gp|U00021|	Mycobacterium leprae cosmid L247. [Mycobacterium leprae]	76	59
198	1	12	356	gp|L13292|	ATPase [Enterococcus hirae] >pir|S|A45995 Cu2+-transporting ATPase (EC 3.6.1.-) - Enterococcus hirae	76	47
6	3	3386	1311	gp|L44894|	excinuclease ABC subunit A [Haemophilus influenzae] >gp|U32711|HIU32711_1 excinuclease ABC subunit A [Haemophilus influenzae] >gp|U00071|HIU00071_64 excinuclease ABC subunit A [Haemophilus influenzae] >gp|U32820|HIU32820_1 excinuclease ATPase subunit [Hae	75	59
8	32	18366	18902	gp|L47971|	ribosomal protein S5 [Bacillus subtilis]	75	54
8	35	21574	21807	gp|M26414|	ribosomal protein S11 [Bacillus subtilis] >gp|L47971|BACRPLP_20 ribosomal protein S11 [Bacillus subtilis] >gp|M13957|BACRPOA_2 B.subtilis DNA sequence of the rpsM-rpoA interval. [Bacillus subtilis] >pir|S|R3BSS1 ribosomal protein S11 - Bacillus subtilis	75	60
15	5	2866	3408	gp|X15981|	E. coli sbcC gene (ORF-45) for SbcC. [Escherichia coli] >pir|S|BVECSC sbcC protein - Escherichia coli >gp|M64787|ECOARAJ_1 sbcC gene product [Escherichia coli] {SUB 378-1048}	75	62
22	23	14859	15518	gp|L45893|	periplasmic serine protease Do and heat shock protein	75	48

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					[Haemophilus influenzae] >gp|U32805|HIU32805_12 periplasmic serine protease Do and heat shock protein [Haemophilus influenzae] >gp|U00082|HIU00082_28 periplasmic serine protease Do and heat shock prote		
24	4	1217	999	gp|J05510|	Rat inositol-1,4,5-triphosphate receptor mRNA, complete cds. [Rattus norvegicus] >pir|S|B36579 inositol 1,4,5-triphosphate receptor 2 - rat >gp|U38665|RNU38665_1 inositol 1,4,5-trisphosphate receptor [Rattus norvegicus] (SUB 1612-1859)	75	58
26	1	266	3	gp|Y00402|	Drosophila melanogaster mRNA for phosphoenolpyruvate carboxykinase (GTP) (PEPCK, EC 4.1.1.32). [Drosophila melanogaster] >pir|S|QYFFGM phosphoenolpyruvate carboxykinase (GTP) (EC 4.1.1.32) precursor, mitochondrial - fruit fly (Drosophila melanogaster)	75	43
28	11	5554	4961	gp|L45197|	putative glucose-6-phosphate dehydrogenase isozyme [Haemophilus influenzae] >gp|U32737|HIU32737_6 putative glucose-6-phosphate dehydrogenase isozyme [Haemophilus influenzae] >gp|U00074|HIU00074_77 putative glucose-6-phosphate dehydrogenase isozyme [Haemop	75	60
28	13	6156	5662	gp|L45199|	glucose-6-phosphate 1-dehydrogenase [Haemophilus influenzae] >gp|U32737|HIU32737_8 glucose-6-phosphate 1-dehydrogenase [Haemophilus influenzae] >gp|U00074|HIU00074_79 glucose-6-phosphate 1-dehydrogenase [Haemophilus influenzae] >gp|U32846|HIU32846_7 gluco	75	56
35	38	19740	21392	gp|D84214|	YbbQ [Bacillus subtilis]	75	51
37	1	799	2	gp|L29053|	htpG gene product [Vibrio fischeri]	75	58
37	3	922	2097	gp|X70943|	aspartyl-tRNA synthetase [Thermus aquaticus thermophilus] >pir|S|S33743 aspartate--tRNA ligase (EC 6.1.1.12) - Thermus aquaticus	75	58
37	16	10484	9618	gp|D64002|	hypothetical protein [Synechocystis sp.]	75	56
38	2	348	701	gp|U14345|	6-phosphogluconate dehydrogenase [Salmonella enterica]	75	60
41	8	2153	2428	gp|U21157|	sarcolemmal associated protein-3 [Oryctolagus cuniculus]	75	65
47	8	4123	4566	gp|L39892|	trxA gene product [Chlamydia psittaci]	75	52

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

48	11	8831	8241	gp|Z32850|	pyrophosphate-dependent phosphofructokinase beta subunit [Ricinus communis]	75	56
54	26	8599	8018	gp|M60917|	purine nucleoside phosphorylase [Escherichia coli] >gp|U14003|ECOUW93_295 purine-nucleoside phosphorylase [Escherichia coli] >pir|S|A27854 purine-nucleoside phosphorylase (EC 2.4.2.1) - Escherichia coli	75	50
64	25	10848	11318	gp|X58114|	testis-specific RNA [Drosophila hydei]	75	62
91	2	165	1025	gp|M88581|	transfer RNA-Leu synthetase [Bacillus subtilis] >pir|S|A41882 leucine--tRNA ligase (EC 6.1.1.4) - Bacillus subtilis	75	66
113	3	1967	2683	gp|X64705|	secA gene product [Antithamnion sp.] >pir|S|S42707 secA protein - red alga (Antithamnion sp.)	75	60
137	11	4611	5657	gp|L44711|	DNA mismatch repair protein [Haemophilus influenzae] >gp|U32692|HIU32692_3 DNA mismatch repair protein [Haemophilus influenzae] >gp|U00069|HIU00069_65 DNA mismatch repair protein [Haemophilus influenzae] >gp|U32801|HIU32801_3 DNA mismatch repair protein [	75	61
228	1	20	397	gp|U03991|	beta-galactosidase alpha peptide [Cloning vector pUC1918] >gp|U33186|CVU33186_1 beta-galactosidase alpha peptide [Cloning vector pSUM36] {SUB 41-96}	75	72
245	1	251	469	gp|Z54206|	BTIF [Bovine herpesvirus 1]	75	66
420	1	81	236	gp|L23651|	C. elegans cosmid C29E4. [Caenorhabditis elegans] >pir|S|S44767 C29E4.1 protein - Caenorhabditis elegans	75	75
458	1	184	2	gp|L06147|	golgin-95 [Homo sapiens] >pir|S|JH0821 95K golgi antigen - human	75	67
8	24	13252	13551	gp|X54994|	ribosomal protein S19 [Bacillus stearothermophilus] >pir|S|R3BS19 ribosomal protein S19 - Bacillus stearothermophilus	74	56
8	33	18841	19551	gp|L47971|	ribosomal protein L15 [Bacillus subtilis] >gp|D00619|BACSECY_2 B.subtilis secY gene. [Bacillus subtilis] >gp|M31102|BACSPCR_3 B.subtilis spectinomycin resistance (spc) genes, complete cds. [Bacillus subtilis] >gp|M31102|BACSPCR_3 Bacillus subtilis spectin	74	55

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

18	7	6261	5623	gp|X68709|	whiG-Stv gene product [Streptoverticillium griseocarneum] >pir|S|S29615 whiG protein - Streptoverticillium griseocarneum	74	55
21	16	10638	10162	gp|U18997|	Escherichia coli K-12 chromosomal region from 67.4 to 76.0 minutes. [Escherichia coli]	74	54
26	25	14169	14657	gp|U39691|	methylgalactoside permease ATP-binding protein [Mycoplasma genitalium] >pir|S|B64213 methylgalactoside permease ATP-binding protein homolog - Mycoplasma genitalium (SGC3) >gp|U02149|MGU02149_1 Mycoplasma genitalium random genomic clone sc8a, partial cds.	74	52
29	16	6788	6084	gp|U00013|	pps1 [Mycobacterium leprae]	74	46
29	17	7660	6170	gp|U00013|	pps1 [Mycobacterium leprae]	74	54
41	5	1572	1955	gp|L15191|	phosphoenolpyruvate:sugar phosphotransferase system enzyme I [Streptococcus mutans] >gp|L15191|STRPHOSPHO_2 phosphoenolpyruvate:sugar phosphotransferase system enzyme I [Streptococcus mutans]	74	60
70	2	856	1119	gp|X79146|	lmbS gene product [Streptomyces lincolnensis] >pir|S|S44965 lmbS protein - Streptomyces lincolnensis	74	56
79	2	893	1453	gp|L76303|	flagellar basal body rod protein [Borrelia burgdorferi]	74	50
80	1	325	2	gp|M80473|	uvr/dinA gene product [Bacillus subtilis] >gp|M64048|BACDINA76_2 Bacillus subtilis DNase inhibitor (dinA76) gene, complete cds and promoter region. [Bacillus subtilis] {SUB 1-57}	74	58
82	4	2176	1022	gp|S56812|	phospho enol pyruvate carboxykinase...F-ATPase epsilon subunit [Chlorobium limicola, Genomic, 4 genes, 5477 nt]. [Chlorobium limicola]	74	64
100	18	4634	4086	gp|L45070|	glutamine amidotransferase [Haemophilus influenzae] >gp|U32726|HIU32726_10 glutamine amidotransferase [Haemophilus influenzae] >gp|U00073|HIU00073_51 glutamine amidotransferase [Haemophilus influenzae] >gp|U32835|HIU32835_5 glutamine amidotransferase [Hae	74	58
101	6	1817	2752	gp|J03885|	K.pneumoniae oxalacetate decarboxylase alpha subunit gene, complete cds. [Klebsiella pneumoniae] >pir|S|A28088 oxaloacetate	74	59

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					decarboxylase (EC 4.1.1.3) alpha chain - Klebsiella pneumoniae		
101	19	9164	10120	gp|U00024|	u0002d [Mycobacterium tuberculosis]	74	53
117	1	913	305	gp|L18867|	ATP-dependent protease ATPase subunit [Escherichia coli] >pir|S|A48709 ATP-dependent Clp proteinase (EC 3.4.-.-) regulatory chain X - Escherichia coli	74	52
141	6	3736	4149	gp|U18997|	Escherichia coli K-12 chromosomal region from 67.4 to 76.0 minutes. [Escherichia coli]	74	52
159	2	1170	1460	gp|X63757|	spoIIAA gene product [Bacillus megaterium] >pir|S|A48402 stage II sporulation protein spoIIAA - Bacillus megaterium	74	44
604	1	2	844	gp|U28377|	metK gene product [Escherichia coli]	74	60
4	4	3687	3995	gp|U44118|	ribosomal protein L20 [Pseudomonas syringae pv. syringae]	73	44
8	23	12415	13254	gp|Z21677|	ribosomal protein L2 [Thermotoga maritima] >pir|S|S40191 ribosomal protein L2 - Thermotoga maritima	73	63
11	25	12075	13619	gp|Z18631|	ORF2 gene product [Bacillus subtilis] >pir|S|C36905 nusA homolog - Bacillus subtilis	73	50
19	11	8644	9660	gp|M63655|	delta-2-isopentenyl pyrophosphate transferase [Escherichia coli] >gp|U14003|ECOUW93_83 tRNA delta-2-isopentenylpyrophosphate (IPP) transferase [Escherichia coli] >pir|S|B37318 delta(2)-isopentenylpyrophosphate transferase (EC 2.5.1.-) - Escherichia coli >	73	48
22	5	4833	3916	gp|U18997|	Escherichia coli K-12 chromosomal region from 67.4 to 76.0 minutes. [Escherichia coli]	73	52
23	2	32	616	gp|X74357|	skp gene product [Pasteurella multocida] >pir|S|S47341 skp protein - Pasteurella multocida	73	42
26	4	2728	3381	gp|L45034|	hypothetical protein (SP:P31216) [Haemophilus influenzae] >gp|U32723|HIU32723_4 hypothetical protein (SP:P31216) [Haemophilus influenzae] >gp|U00073|HIU00073_15 hypothetical protein (SP:P31216) [Haemophilus influenzae] >gp|U32832|HIU32832_2 GTPase [Haemop	73	57
26	7	4429	4824	gp|D26185|	3'-exo-deoxyribonuclease [Bacillus subtilis]	73	57
28	2	1598	846	gp|X59551|	ORF246 gene product [Escherichia coli] >gp|D10165|ECORUVC_3 Orf26 [Escherichia coli] >pir|S|C38113	73	52

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					26K hypothetical protein (ruvC 5' region) - Escherichia coli		
30	10	4632	5738	gp|U17591|	primary sigma factor [Borrelia burgdorferi]	73	53
34	3	1318	2271	gp|X82072|	orf2 gene product [Pseudomonas aeruginosa] >pir|S|S49379 hypothetical protein 2 - Pseudomonas aeruginosa >pir|S|S57900 hypothetical protein 2 - Pseudomonas aeruginosa	73	53
38	1	2	571	gp|L45194|	6-phosphogluconate dehydrogenase, decarboxylating [Haemophilus influenzae] >gp|U32737|HIU32737_3 6-phosphogluconate dehydrogenase, decarboxylating [Haemophilus influenzae] >gp|U00074|HIU00074_74 6-phosphogluconate dehydrogenase, decarboxylating [Haemophil	73	59
42	15	8010	9143	gp|Z11839|	ribosomal protein L1 [Thermotoga maritima] >pir|S|R5HG1T ribosomal protein L1 - Thermotoga maritima	73	49
51	3	4067	4717	gp|L25288|	gyrase-like protein alpha subunit [Staphylococcus aureus]	73	54
56	6	2887	3738	gp|D64116|	ORF4 [Bacillus subtilis]	73	50
57	4	2201	3697	gp|X79516|	membrane ATPase [Haloferax volcanii] >pir|S|S45145 H+-transporting ATP synthase (EC 3.6.1.34) beta chain - Haloferax volcanii >pir|S|S55896 membrane ATPase B chain - Haloferax volcanii	73	60
57	7	5549	5352	gp|U07825|	histone variant H1.1(a) [Parechinus angulosus]	73	31
79	3	1482	1985	gp|U43739|	FlgC [Borrelia burgdorferi] >gp|L76303|BORFTSA_8 flagellar basal body rod protein [Borrelia burgdorferi]	73	57
100	30	12237	11806	gp|D64003|	hypothetical protein [Synechocystis sp.] >gp|D64003|SYCSLLE_14 hypothetical protein [Synechocystis sp.]	73	50
142	11	13952	13527	gp|X75627|	spoIIIE gene product [Coxiella burnetii] >pir|S|S43132 spoIIIE protein - Coxiella burnetii >pir|S|S31759 hypothetical protein 274 - Coxiella burnetii {SUB 505-778}	73	55
167	4	1451	2062	gp|X86481|	rrf gene product [Clostridium perfringens]	73	50
668	2	143	373	pir|S|S33716	proline dehydrogenase - Salmonella typhimurium	73	60
2	11	7849	9303	gp|L31845|	UDP-N-acetyl muramate-alanine ligase [Bacillus subtilis]	72	45
5	6	3790	4449	gp|M25899|	antimicrobial protein [Artificial gene]	72	38
6	7	5180	4983	gp|X72857|	traX gene product [Streptomyces coelicolor] >pir|S|S39854	72	63

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					hypothetical protein X - Streptomyces coelicolor >pir|S|S32232 traX protein - Streptomyces coelicolor		
8	29	16835	16990	gp|Z28203|	TOR2 gene product [Saccharomyces cerevisiae] >pir|S|S38040 TOR2 protein - yeast (Saccharomyces cerevisiae)	72	55
16	14	10815	9463	gp|D64005|	hypothetical protein [Synechocystis sp.]	72	55
18	5	5680	5177	gp|X68709|	whiG-Stv gene product [Streptoverticillium griseocarneum] >pir|S|S29615 whiG protein - Streptoverticillium griseocarneum	72	46
25	4	1968	1804	gp|M22039|	fructose-bisphosphate aldolase [Bacillus subtilis] >gp|Z49782|BSDNA320D_13 fructose biphosphate aldolase [Bacillus subtilis] >pir|S|D32354 fructose-bisphosphate aldolase (EC 4.1.2.13) - Bacillus subtilis >gp|S42590|S42590_1 fructose-1,6-bisphosphate aldol	72	51
26	27	15501	16058	gp|U39691|	ribose transport system permease protein [Mycoplasma genitalium] >pir|S|C64213 ribose transport system permease homolog - Mycoplasma genitalium (SGC3)	72	48
27	3	1873	956	gp|Z49633|	S.cerevisiae chromosome X reading frame ORF YJR133w. [Saccharomyces cerevisiae] >pir|S|S57156 hypothetical protein YJR133w - yeast (Saccharomyces cerevisiae)	72	48
35	8	4530	5426	gp|M14339|	S.pneumoniae DpnII gene region encoding dpnM, dpnA, dpnB, complete cds. [Streptococcus pneumoniae] >gp|M11226|STRDPNM_1 S.pneumoniae DpnM gene encoding Dpn II DNA methylase, complete cds. [Streptococcus pneumoniae] >gp|M14339|STRDPN2A_2 dpnM gene product	72	59
35	31	17563	18165	gp|Z11165|	641 aa (68 kD) gene product of ORF641 [Rhodobacter capsulatus] >pir|S|G28771 hypothetical protein C2814 (photosynthetic gene cluster) - Rhodobacter capsulatus	72	57
36	15	9228	10505	gp|U39703|	glycyl-tRNA synthetase [Mycoplasma genitalium] >pir|S|G64227 glycine--tRNA ligase (EC 6.1.1.14) - Mycoplasma genitalium (SGC3)	72	51
62	9	3780	4169	gp|M64519|	transport protein [Escherichia coli] >pir|S|A40840 spermidine/putrescine transport protein A - Escherichia coli	72	62
69	11	6417	5293	gp|X04027|	E. coli mutD(dnaQ)-rnh region for DNA polymerase III epsilon	72	63

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					subunit and RNAase H. [Escherichia coli] >gp|K00552|ECORNH_1 E.coli rnh gene coding for ribonuclease H. [Escherichia coli] >gp|K00985|ECORNHQ_2 ribonuclease H [Escherichia coli] >gp|V00337|ECRNH		
79	12	6913	7581	gp|M72718|	B.subtilis flaA locus operon. [Bacillus subtilis] >gp|X56049|BSFLAAO_3 B. subtilis FlaA locus operon. [Bacillus subtilis] >pir|S|PWBSAS H+-transporting ATP synthase alpha chain homolog - Bacillus subtilis	72	53
82	2	475	281	gp|Z15025|	Bat2 gene product [Homo sapiens] >pir|S|S37671 bat2 protein - human	72	72
82	3	1136	480	gp|Y00402|	Drosophila melanogaster mRNA for phosphoenolpyruvate carboxykinase (GTP) (PEPCK, EC 4.1.1.32). [Drosophila melanogaster] >pir|S|QYFFGM phosphoenolpyruvate carboxykinase (GTP) (EC 4.1.1.32) precursor, mitochondrial - fruit fly (Drosophila melanogaster)	72	57
89	1	1041	4	gp|M63176|	helicase [Staphylococcus aureus] >pir|S|S27667 DNA helicase pcrA - Staphylococcus aureus >pir|S|S39923 pcrA protein - Staphylococcus aureus	72	61
89	14	12410	11475	gp|U12513|	PPi-dependent phosphofructo-1-kinase [Entamoeba histolytica] >gp|U12513|EHU12513_1 PPi-dependent phosphofructo-1-kinase [Entamoeba histolytica] >pir|S|S52082 PPi-dependent phosphofructo-1-kinase - Entamoeba histolytica	72	49
101	4	1838	1557	gp|D25220|	selenoprotein P like protein [Bos taurus] >gp|D25220|BOVSPP_1 selenoprotein P like protein [Bos taurus]	72	44
105	4	2929	1649	pir|S|SYBS YF	tyrosine--tRNA ligase (EC 6.1.1.1) - Bacillus stearothermophilus	72	54
116	5	930	1568	gp|X81990|	leader peptidase I [Phormidium laminosum] >pir|S|S51921 leader peptidase (EC 3.4.99.36) - oscillatoriacean cyanobacterium	72	50
134	2	1073	477	gp|U33210|	methyl accepting chemotaxis homolog [Treponema denticola]	72	54
136	1	2	268	gp|U15140|	ribosomal protein IF-1 [Mycobacterium bovis]	72	52
144	2	119	310	gp|L28810|	regulatory protein [Aspergillus nidulans]	72	61
161	2	787	464	gp|M58002|	bacterial cell wall hydrolase [Streptococcus faecalis] >pir|S|A38109	72	39

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					autolysin - Enterococcus faecalis		
186	2	229	603	gp|D64000|	hypothetical protein [Synechocystis sp.]	72	59
555	2	157	456	gp|X77515|	pyruvate oxidoreductase [Rhodospirillum rubrum] >pir|S|S41961 pyruvate oxidoreductase - Rhodospirillum rubrum	72	49
5	2	858	1499	gp|M11330|	CDP-diglyceride synthetase [Escherichia coli] >gp|D83536|ECOTSF_5 phosphatidate cytidylyltransferase [Escherichia coli] >pir|S|SYECDG phosphatidate cytidylyltransferase (EC 2.7.7.41) - Escherichia coli	71	52
11	28	14189	16321	gp|X00513|	E.coli nusA operon including genes for Met-tRNA-f2 (metY), 15 kd protein, NusA protein (nusA), and initiation factor IF2 (infB). [Escherichia coli] >gp|U18997|ECOUW67_98 protein chain initiation factor 2 [Escherichia coli]	71	54
13	8	5270	3489	gp|L45934|	hypothetical protein (GB:U14003_302) [Haemophilus influenzae] >gp|U32809|HIU32809_12 hypothetical protein (GB:U14003_302) [Haemophilus influenzae] >gp|U00082|HIU00082_66 hypothetical protein (GB:U14003_302) [Haemophilus influenzae] >gp|U32755|HIU32755_8 t	71	50
16	35	21358	20309	gp|M86351|	triacylglycerol acylhydrolase [Streptomyces sp.] >pir|S|JN0490 28K lipase precursor - Streptomyces sp. (strain M11)	71	53
21	7	5991	4771	gp|D16437|	PacS [Synechococcus sp.] >pir|S|S36741 cation-transporting ATPase pacS - Synechococcus sp.	71	49
29	11	4929	4390	gp|D64004|	hypothetical protein [Synechocystis sp.] >gp|D64004|SYCSLRF_7 hypothetical protein [Synechocystis sp.]	71	43
35	37	18671	19465	gp|L46071|	hypothetical protein (SP:P26242) [Haemophilus influenzae] >gp|U32822|HIU32822_12 hypothetical protein (SP:P26242) [Haemophilus influenzae] >gp|U00084|HIU00084_34 hypothetical protein (SP:P26242) [Haemophilus influenzae] >gp|U32768|HIU32768_15 transketolas	71	56
35	40	24418	23495	gp|U51032|	D9651.10 gene product [Saccharomyces cerevisiae]	71	41
41	13	5053	5550	gp|X05991|	Drosophila Cs gene. [Drosophila melanogaster] >pir|S|S01103	71	64

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					hypothetical protein 2 - fruit fly (Drosophila melanogaster)		
42	11	5125	5469	gp|L46067|	H. influenzae predicted coding region HI1435 [Haemophilus influenzae] >gp|U32822|HIU32822_8 H. influenzae predicted coding region HI1435 [Haemophilus influenzae] >gp|U00084|HIU00084_30 H. influenzae predicted coding region HI1435 [Haemophilus influenzae]	71	50
42	24	16051	17931	gp|U00006|	DNA-directed RNA polymerase, beta'-subunit [Escherichia coli] >gp|X04642|STRPOB_2 S. typhimurium rpoB gene for RNA polymerase beta subunit. [Salmonella typhimurium] {SUB 1-20}	71	53
47	21	12695	12396	gp|X74933|	tRNA(m1G37)methyltransferase [Salmonella typhimurium] >pir|S|S37175 tRNA (guanine-N1-)-methyltransferase (EC 2.1.1.31) - Salmonella typhimurium	71	40
48	4	2455	2279	gp|Z29115|	T26G10.1 [Caenorhabditis elegans] >pir|S|S40731 hypothetical protein - Caenorhabditis elegans	71	52
48	7	5241	4102	gp|M91598|	3-phosphoglycerate kinase [Yarrowia lipolytica]	71	46
52	5	6567	4714	gp|X73141|	hemolysin [Serpulina hyodysenteriae]	71	48
57	6	6423	4474	gp|L45709|	hypothetical protein (SP:P37024) [Haemophilus influenzae] >gp|U32787|HIU32787_7 hypothetical protein (SP:P37024) [Haemophilus influenzae] >gp|U00080|HIU00080_36 hypothetical protein (SP:P37024) [Haemophilus influenzae] >pir|S|G64165 hypothetical protein H	71	54
61	8	3121	3765	gp|L45095|	hypothetical protein (GB:D26185_102) [Haemophilus influenzae] >gp|U32728|HIU32728_13 hypothetical protein (GB:D26185_102) [Haemophilus influenzae] >gp|U00073|HIU00073_76 hypothetical protein (GB:D26185_102) [Haemophilus influenzae] >gp|U32837|HIU32837_6 H	71	44
63	19	7365	7027	gp|L45782|	hypothetical protein (SP:P33995) [Haemophilus influenzae] >gp|U32794|HIU32794_8 hypothetical protein (SP:P33995) [Haemophilus influenzae] >gp|U00081|HIU00081_13 hypothetical protein (SP:P33995) [Haemophilus influenzae] >gp|U32740|HIU32740_8 ATP(GTP)-utili	71	53
63	21	9982	8696	gp|M19488|	S.typhimurium prsA gene encoding phosphoribosylpyrophosphate	71	49

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					synthetase, complete cds. [Salmonella typhimurium] >pir|S|KIEBRT ribose-phosphate pyrophosphokinase (EC 2.7.6.1) - Salmonella typhimurium		
65	5	1824	2435	gp|D55650|	adenylate cyclase [Anabaena cylindrica]	71	50
118	1	650	1324	gp|Z36878|	putative nicotinate phosphorybosyltranferase [Saccharomyces cerevisiae] >pir|S|S51845 probable nicotinate phosphoribosyltransferase (EC 2.4.2.11) - yeast (Saccharomyces cerevisiae) >gp|L11274|YSCNPTRPBX_1 nicotinate phosphoribosyltransferase [Saccharomyce	71	59
131	2	1615	746	gp|M30942|	transfer RNA-Ile synthetase [Tetrahymena thermophila] >pir|S|A42399 isoleucyl-tRNA synthetase, ileRS - Tetrahymena thermophila (SGC5)	71	52
150	9	8131	7304	gp|D26185|	regulation of Spo0J and Orf283 (probable) [Bacillus subtilis] >gp|X62539|BSORIGS_10 B.subtilis genes rpmH, rnpA, 50kd, gidA and gidB. [Bacillus subtilis] >pir|S|S18080 hypothetical protein 5 - Bacillus subtilis	71	55
157	3	1725	1366	gp|L44678|	hypothetical protein (SP:P05848) [Haemophilus influenzae] >gp|U32688|HIU32688_13 hypothetical protein (SP:P05848) [Haemophilus influenzae] >gp|U00069|HIU00069_32 hypothetical protein (SP:P05848) [Haemophilus influenzae] >gp|U32798|HIU32798_1 H. influenzae	71	50
2	13	13010	10815	gp|D26185|	transcription-repair coupling factor [Bacillus subtilis]	70	50
4	3	2796	3344	gp|U32757|	initiation factor 3 [Haemophilus influenzae] >gp|L45952|HEAHI1318_1 initiation factor 3 [Haemophilus influenzae] {SUB 49-183}	70	46
8	21	10812	11444	gp|X67014|	ribosomal protein L3 [Bacillus stearothermophilus] >pir|S|S24363 ribosomal protein L3 - Bacillus stearothermophilus	70	46
16	8	5949	5515	gp|X73141|	hemolysin [Serpulina hyodysenteriae]	70	36
23	7	3492	5213	gp|Z38002|	serine hydroxymethyltransferase [Bacillus subtilis] >pir|S|S49363 serine hydroxymethyltransferase - Bacillus subtilis	70	49
29	28	13092	13259	gp|X05790|	alpha-galactosidase [Homo sapiens] >gp|X14448|HSGLA_1 alpha-D-galactosidase A [Homo sapiens] >pir|S|GBHUA alpha-	70	50

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					galactosidase (EC 3.2.1.22) A precursor - human >gp|M13571|HUMAGALAA_1 GLA gene product [Homo sapiens] (SUB 27-429)		
35	1	306	4	gp|U23514|	F48E8.6 gene product [Caenorhabditis elegans]	70	50
35	6	3610	3909	gp|M76547|	acyl carrier protein [Chloroplast Cryptomonas phi] >pir|S|C41609 acyl carrier protein acpA - Cryptomonas sp. chloroplast (strain Phi) (SUB 2-81)	70	45
35	36	18773	18552	gp|X87331|	ORF OR26.23 gene product [Saccharomyces cerevisiae] >gp|S69545|S69545_1 Dhs1 [Saccharomyces cerevisiae] (SUB 220-702)	70	35
37	6	1980	2591	gp|X70943|	aspartyl-tRNA synthetase [Thermus aquaticus thermophilus] >pir|S|S33743 aspartate--tRNA ligase (EC 6.1.1.12) - Thermus aquaticus	70	53
39	8	3707	3195	gp|U32699|	acid phosphatase (?) [Haemophilus influenzae]	70	47
44	5	1398	2432	gp|X73140|	hemolysin [Serpulina hyodysenteriae]	70	49
52	3	4080	2803	gp|U10400|	YHR011w gene product [Saccharomyces cerevisiae] >pir|S|S46786 serine--tRNA ligase homolog - yeast (Saccharomyces cerevisiae)	70	46
56	9	3984	5726	gp|D64116|	ORF4 [Bacillus subtilis]	70	48
62	8	3523	4071	gp|L45980|	spermidine/putrescine transport ATP-binding protein [Haemophilus influenzae] >gp|U32813|HIU32813_12 spermidine/putrescine transport ATP-binding protein [Haemophilus influenzae] >gp|U00083|HIU00083_23 spermidine/putrescine transport ATP-binding protein [Ha	70	48
74	5	1229	1444	gp|U41536|	F56E3.3 gene product [Caenorhabditis elegans]	70	35
77	12	6229	5666	gp|X03038|	E. coli adk gene for adenylate kinase. [Escherichia coli] >gp|M38777|ECOAPTADK_6 E.coli sequence of the apt-adk region. [Escherichia coli] >pir|S|KIECA adenylate kinase (EC 2.7.4.3) - Escherichia coli >gp|D90259|ECOADKVIS_1 E.coli adk, visA genes and ORFs	70	47
81	14	13662	13255	pir|S|R3BS9	ribosomal protein S9 - Bacillus stearothermophilus	70	54
100	12	3525	2704	gp|L45070|	glutamine amidotransferase [Haemophilus influenzae]	70	54

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					>gp|U32726|HIU32726_10 glutamine amidotransferase [Haemophilus influenzae] >gp|U00073|HIU00073_51 glutamine amidotransferase [Haemophilus influenzae] >gp|U32835|HIU32835_5 glutamine amidotransferase [Hae		
100	19	5203	4832	gp|U14003|	Escherichia coli K-12 chromosomal region from 92.8 to 00.1 minutes. [Escherichia coli] >pir|S|S56432 hypothetical protein o259a - Escherichia coli >pir|S|S46294 peptidylprolyl isomerase (EC 5.2.1.8) - Escherichia coli {SUB 55-259}	70	56
100	29	11866	11027	gp|D64003|	hypothetical protein [Synechocystis sp.] >gp|D64003|SYCSLLE_14 hypothetical protein [Synechocystis sp.]	70	56
101	15	7124	6945	gp|X63765|	ribosomal protein RL9 [Synechococcus sp.] >pir|S|S22206 ribosomal protein L9 - Synechococcus sp.	70	50
121	22	10796	12685	gp|X94607|	Mef1 protein [Saccharomyces cerevisiae]	70	56
123	3	2255	1128	gp|M80215|	uvs402 protein [Streptococcus pneumoniae] >pir|S|A42385 uvr-402 protein - Streptococcus pneumoniae plasmid pSB470	70	51
143	3	3370	872	gp|L45958|	lon protease [Haemophilus influenzae] >gp|U32812|HIU32812_1 lon protease [Haemophilus influenzae] >gp|U00083|HIU00083_2 lon protease [Haemophilus influenzae] >gp|U32757|HIU32757_9 Lon/Sms-related endopeptidase (no ATPase domain) [Haemophilus influenzae] >	70	54
175	2	852	118	gp|L45009|	protein E [Haemophilus influenzae] >gp|U32721|HIU32721_3 protein E [Haemophilus influenzae] >gp|U00072|HIU00072_93 protein E [Haemophilus influenzae] >gp|U32830|HIU32830_3 essential protein [Haemophilus influenzae] >pir|S|H64063 protein E (gpcE) homolog -	70	50
313	1	474	151	gp|D10388|	N-acetylmuramoyl-L-alanine amidase [Bacillus subtilis] >gp|M81324|BACCWLB_1 N-acetylmuramoyl-L-alanine amidase [Bacillus subtilis] >gp|M87645|BACLYTABCD_4 amidase [Bacillus subtilis] >pir|S|B41322 N-acetylmuramoyl-L-alanine amidase (EC 3.5.1.28) 50K precu	70	44
441	1	2	232	gp|U14333|	Arnt [Mus musculus]	70	47
590	1	3	338	gp|M27713|	D.discoideum Thy1 gene, complete cds. [Dictyostelium	70	56

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					discoideum] >pirlS|YXDOTC thymidylate synthase-complementing protein - slime mold (Dictyostelium discoideum)		
641	1	51	575	gp|L34879|	NifS gene product [Anabaena azollae] >gp|L34879|ANAAZNIF_3 NifS gene product [Anabaena azollae]	70	65
4	13	8224	9780	gp|X54548|	serine protease [Salmonella typhimurium] >pirlS|S15337 heat shock protein htrA - Salmonella typhimurium	69	55
13	6	2524	3492	gp|L46000|	threonyl-tRNA synthetase [Haemophilus influenzae] >gp|U32816|HIU32816_6 threonyl-tRNA synthetase [Haemophilus influenzae] >gp|U00083|HIU00083_43 threonyl-tRNA synthetase [Haemophilus influenzae] >gp|U32762|HIU32762_4 threonyl-tRNA synthetase [Haemophilus	69	48
16	13	9323	7899	gp|D64005|	hypothetical protein [Synechocystis sp.]	69	57
16	50	29876	29175	gp|X52094|	flgG protein product (AA 1-260) [Salmonella typhimurium] >pirlS|XMEBFG basal body rod protein flgG - Salmonella typhimurium	69	53
16	51	30892	29936	gp|X52094|	flgG protein product (AA 1-260) [Salmonella typhimurium] >pirlS|XMEBFG basal body rod protein flgG - Salmonella typhimurium	69	54
16	56	34171	33398	gp|D64006|	hypothetical protein [Synechocystis sp.]	69	47
17	6	3294	2881	gp|L09228|	Bacillus subtilis spoVA to serA region. [Bacillus subtilis] >pirlS|S45550 hypothetical protein X8 - Bacillus subtilis	69	47
28	8	4806	3685	gp|L17320|	acetate kinase [Bacillus subtilis] >pirlS|B49935 acetate kinase (EC 2.7.2.1) - Bacillus subtilis	69	47
29	7	3371	2706	gp|D64002|	hypothetical protein [Synechocystis sp.]	69	43
29	19	8526	7603	gp|D64004|	hypothetical protein [Synechocystis sp.] >gp|D64004|SYCSLRF_5 hypothetical protein [Synechocystis sp.]	69	48
30	5	2460	2894	gp|U13165|	DNA primase [Listeria monocytogenes]	69	41
44	2	130	1539	gp|X73140|	hemolysin [Serpulina hyodysenteriae]	69	51
44	9	5167	4751	gp|L25421|	phosphatase [Treponema denticola]	69	50
45	7	7818	8294	gp|V01316|	tRNA synthetase [Saccharomyces cerevisiae]	69	55

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					>gp|J01339|YSCMES1_1 Yeast (S.cerevisiae) methionyl-tRNA synthetase (mes1) gene, complete cds. [Saccharomyces cerevisiae] >gp|J01339|YSCMES1_1 S.cerevisiae methionyl-tRNA synthetase (mes1) gene, complete cds. [Sa		
47	2	1724	822	gp|U19610|	NADH oxidase [Serpulina hyodysenteriae]	69	48
60	4	2516	1218	gp|J04836|	M.barkeri ATPase alpha and beta subunit (atpA and atpB) genes, complete cds. [Methanosarcina barkeri] >pir|S|B34283 H+-transporting ATP synthase (EC 3.6.1.34) beta chain - Methanosarcina barkeri	69	51
67	1	1916	180	pir|S|S09411	spoIIIE protein - Bacillus subtilis >gp|M17445|BACSPIIIA_3 B.subtilis sporulation protein spoIIIEA and spoIIIEB genes, complete cds and open reading frame X, 3' end. [Bacillus subtilis] {SUB 536-787}	69	49
70	7	7470	5644	gp|X52227|	E.coli fhlA gene for the transcriptional activator of the formate hydrogenlyase. [Escherichia coli] >gp|U29579|ECU29579_28 transcriptional activator of the formate hydrogenlyase system [Escherichia coli] >pir|S|S12079 transcriptional activator fhlA - Esch	69	52
79	4	1967	2356	gp|L76303|	flagellar basal body rod protein [Borrelia burgdorferi]	69	36
80	2	356	1201	gp|X73124|	ipa-52r gene product [Bacillus subtilis] >pir|S|S39707 hypothetical protein - Bacillus subtilis	69	48
81	20	17491	17204	gp|M63176|	helicase [Staphylococcus aureus] >pir|S|S27667 DNA helicase pcrA - Staphylococcus aureus >pir|S|S39923 pcrA protein - Staphylococcus aureus	69	48
86	2	345	809	gp|Z21970|	54CP [Chloroplast Arabidopsis thaliana] >pir|S|S36637 signal recognition particle 54CP protein precursor - Arabidopsis thaliana	69	59
86	3	616	1602	gp|X01818|	E. coli trmD operon and nearby regions. [Escherichia coli] >pir|S|S07178 hypothetical protein, 48K (rpsP 5' region) - Escherichia coli	69	46
98	3	3371	1197	gp|U29668|	polynucleotide phosphorylase [Bacillus subtilis]	69	52
100	15	4224	3529	gp|L10328|	glutamine amidotransferase [Escherichia coli]	69	51
100	28	11096	10338	gp|D64003|	hypothetical protein [Synechocystis sp.]	69	53

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					>gp|D64003|SYCSLLE_14 hypothetical protein [Synechocystis sp.]		
103	5	4051	4554	gp|L13078|	antizyme [Escherichia coli] >pir|S|A48291 ornithine decarboxylase inhibitor - Escherichia coli	69	53
140	2	165	599	gp|X54548|	serine protease [Salmonella typhimurium] >pir|S|S15337 heat shock protein htrA - Salmonella typhimurium	69	53
165	3	584	309	gp|U14003|	Escherichia coli K-12 chromosomal region from 92.8 to 00.1 minutes. [Escherichia coli] >pir|S|S56374 hypothetical protein f342 - Escherichia coli	69	57
165	7	1946	2371	gp|D83536|	hypothetical 23.3 kd protein [Escherichia coli] >gp|D15061|ECORRNHK12_5 ORF217 [Escherichia coli] {SUB 5-217}	69	44
376	2	224	628	gp|U09229|	transcription factor [Rattus norvegicus] >pir|S|B53689 homeotic protein CDP2 - rat (fragment)	69	58
555	3	410	757	gp|D64005|	hypothetical protein [Synechocystis sp.]	69	62
629	1	3	692	gp|M91593|	Mycoplasma mycoides SRPM54 gene, complete cds. [Mycoplasma mycoides] >pir|S|S35480 hypothetical protein 1 - Mycoplasma mycoides (SGC3) >pir|S|S27590 hypothetical protein 1 - Mycoplasma mycoides (SGC3) (fragment) {SUB 2-422}	69	48
1	6	1876	2325	gp|U22817|	RuvB Protein [Thermus aquaticus thermophilus]	68	54
8	27	15955	16290	gp|M81748|	ribosomal protein L24 [Bacillus subtilis] >gp|L47971|BACRPLP_5 ribosomal protein L24 [Bacillus subtilis] >gp|X15664|BSSPC_4 B.subtilis S10/spc operon rpmC, rpsQ, rplN, rplX, rplE, rpsN genes. [Bacillus subtilis] >pir|S|R5BS2B ribosomal protein L24 - Bacil	68	50
8	36	21789	22193	gp|U30821|	alpha subunit of RNA polymerase [Cyanelle Cyanophora paradoxa]	68	57
10	7	5142	4447	gp|D64003|	hypothetical protein [Synechocystis sp.] >gp|D64003|SYCSLLE_79 hypothetical protein [Synechocystis sp.]	68	43
11	8	4625	5395	gp|U39703|	DNA helicase II [Mycoplasma genitalium] >pir|S|I64226 DNA helicase II (mutB1) homolog - Mycoplasma genitalium (SGC3)	68	46

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					>gp|X61517|MGMG08_1 M.genitalium random genomic sequence MG08. [Mycoplasma genitalium] {SUB 277-345}		
12	7	4202	4507	gp|U29580|	Escherichia coli K-12 genome; approximately 62 minute region. [Escherichia coli]	68	34
16	28	16838	16275	gp|X65028|	rotamase [Synechococcus sp. (PCC 7942)] >pir|S|CSYC42 peptidylprolyl isomerase (EC 5.2.1.8) - Synechococcus sp. (PCC 7942)	68	68
21	8	7156	5951	gp|D16437|	PacS [Synechococcus sp.] >pir|S|S36741 cation-transporting ATPase pacS - Synechococcus sp.	68	51
22	20	13872	13486	gp|L46319|	endonuclease III [Haemophilus influenzae] >gp|U32842|HIU32842_1 endonuclease III [Haemophilus influenzae] >gp|U00086|HIU00086_51 endonuclease III [Haemophilus influenzae] >gp|U32788|HIU32788_7 endonuclease III [Haemophilus influenzae] >pir|S|G64136 endonu	68	53
24	2	479	1228	gp|U09189|	loricrin [Mus musculus] >gp|M34398|MUSLRCNA_1 loricrin [Mus musculus] >pir|S|A35628 loricrin - mouse	68	64
28	10	4950	4804	gp|U02025|	insulin-like growth factor binding protein 5 [Mus musculus] >pir|S|A54259 insulin-like growth factor binding protein 5 - mouse (fragment) {SUB 1-111}	68	45
30	4	1681	2442	gp|L45098|	aminodeoxychorismate lyase [Haemophilus influenzae] >gp|U32728|HIU32728_16 aminodeoxychorismate lyase [Haemophilus influenzae] >gp|U00073|HIU00073_79 aminodeoxychorismate lyase [Haemophilus influenzae] >gp|U32837|HIU32837_9 H. influenzae predicted coding	68	52
36	17	10774	11841	gp|U12735|	aminoalcoholphosphotransferase [Glycine max]	68	54
48	10	8283	7114	gp|Z32850|	pyrophosphate-dependent phosphofructokinase beta subunit [Ricinus communis]	68	52
49	13	8601	9068	gp|X64324|	calsequestrin [Rana esculenta] >pir|S|S22418 calsequestrin - edible frog	68	47
61	11	5514	6017	gp|L42554|	pleD gene product [Caulobacter crescentus] >gp|L42554|CCRPLED_1 pleD gene product [Caulobacter crescentus]	68	46

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

66	2	1223	720	gp|M28883|	Rabbit macrophage cationic peptide 1 (MCP-1) mRNA, complete cds. [Oryctolagus cuniculus] >gp|M28072|RABMCP1AA_1 Rabbit macrophage cationic peptide 1 (MCP-1) gene, complete cds. [Oryctolagus cuniculus] >pir|S|A45811 macrophage cationic peptide 1 precursor	68	45
70	28	17845	18918	gp|L23147|	phosphotransacetylase [Methanosarcina thermophila] >pir|S|A49338 phosphate acetyltransferase (EC 2.3.1.8) - Methanosarcina thermophila >gp|U50189|MTU50189_1 phosphate acetyltransferase [Methanosarcina thermophila] {SUB 1-22}	68	52
75	11	4924	4022	gp|M26934|	E.coli ansA-ORF1 gene pair, complete cds. [Escherichia coli] >pir|S|QQECA5 hypothetical 23K protein (ansA 3' region) - Escherichia coli	68	50
81	15	14103	13660	gp|L46075|	ribosomal protein L13 [Haemophilus influenzae] >gp|U32823|HIU32823_2 ribosomal protein L13 [Haemophilus influenzae] >gp|U00084|HIU00084_38 ribosomal protein L13 [Haemophilus influenzae] >gp|U32769|HIU32769_2 ribosomal protein L13 [Haemophilus influenzae]	68	48
104	1	99	389	gp|D26185|	unknown [Bacillus subtilis]	68	31
121	3	1239	772	gp|L45263|	polypeptide deformylase (formylmethionine deformylase) [Haemophilus influenzae] >gp|U32745|HIU32745_1 polypeptide deformylase (formylmethionine deformylase) [Haemophilus influenzae] >gp|U00075|HIU00075_42 polypeptide deformylase (formylmethionine deformyl	68	52
135	1	128	478	gp|L44884|	protein-export membrane protein [Haemophilus influenzae] >gp|U32710|HIU32710_5 protein-export membrane protein [Haemophilus influenzae] >gp|U00071|HIU00071_55 protein-export membrane protein [Haemophilus influenzae] >gp|U32819|HIU32819_4 protein-export me	68	45
136	2	1040	387	gp|D64004|	hypothetical protein [Synechocystis sp.] >gp|D64004|SYCSLRF_28 hypothetical protein [Synechocystis sp.]	68	59
142	6	8745	9422	gp|L45084|	recombinational DNA repair protein [Haemophilus influenzae] >gp|U32728|HIU32728_2 recombinational DNA repair protein	68	47

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					[Haemophilus influenzae] >gp|U00073|HIU00073_65 recombinational DNA repair protein [Haemophilus influenzae] >gp|U32836|HIU32836_4 recombin		
150	14	12885	11392	gp|L37094|	protease [Bartonella bacilliformis]	68	48
170	2	1730	1008	gp|U43739|	FtsZ [Borrelia burgdorferi]	68	44
186	1	2	265	gp|L44885|	protein-export membrane protein [Haemophilus influenzae] >gp|U32710|HIU32710_6 protein-export membrane protein [Haemophilus influenzae] >gp|U00071|HIU00071_56 protein-export membrane protein [Haemophilus influenzae] >gp|U32819|HIU32819_5 protein-export me	68	36
323	1	184	468	gp|V00328|	recA gene product [Escherichia coli] >pir|S|RQECA recA protein - Escherichia coli {SUB 2-353}	68	47
5	7	4447	6303	gp|S63246|	CHL15 gene product [Saccharomyces cerevisiae]	67	32
13	4	1692	2564	gp|V00291|	E.coli thrS, infC, rplT, pheS, pheT and himA genes encoding threonyl-tRNA synthetase, initiation factor IF3, ribosomal protein L20, phenylalanyl-tRNA synthetase and the alpha-subunit of the host integration factor. [Escherichia coli] >pir|S|SYECTT threoni	67	47
19	21	16547	15312	pir|S|XYEBET	protein-glutamate methylesterase (EC 3.1.1.61) - Salmonella typhimurium >pir|S|A26119 protein-glutamate O-methyltransferase (EC 2.1.1.80) - Salmonella typhimurium (fragment) {SUB 277-306}	67	50
23	3	624	1487	gp|M63007|	mutS gene product [Azotobacter vinelandii] >pir|S|A53296 DNA mismatch repair protein MutS - Azotobacter vinelandii	67	41
23	8	6224	5187	gp|U39732|	cell division protein [Mycoplasma genitalium] >pir|S|E64250 cell division protein ftsH - Mycoplasma genitalium (SGC3)	67	50
26	28	16025	16981	gp|Z33126|	membrane forming protein [Mycoplasma capricolum] >pir|S|S48611 hypothetical protein - Mycoplasma capricolum (SGC3) (fragment) {SUB 1-101}	67	32
28	14	7151	6051	gp|L45199|	glucose-6-phosphate 1-dehydrogenase [Haemophilus influenzae] >gp|U32737|HIU32737_8 glucose-6-phosphate 1-dehydrogenase [Haemophilus influenzae] >gp|U00074|HIU00074_79 glucose-6-phosphate 1-dehydrogenase [Haemophilus influenzae]	67	50

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					>gp|U32846|HIU32846_7 gluco		
30	6	2930	3220	gp|M10040|	B.subtilis dnaE gene encoding DNA primase, complete cds. [Bacillus subtilis] >gp|X03897|BSSIG43_2 Bacillus subtilis sigma 43 operon with P23-dnaE-rpoD genes (dnaE for DNA primase, rpoD for RNA polymerase). [Bacillus subtilis] >pir|S|A22282 primase - Bacil	67	42
35	15	8921	9520	gp|J03762|	E.coli thioredoxin reductase gene, complete cds. [Escherichia coli] >pir|S|RDECT thioredoxin reductase (NADPH) (EC 1.6.4.5) - Escherichia coli >gp|L21749|ECOCYDD_1 thioredoxin reductase [Escherichia coli] {SUB 244-321}	67	52
36	3	2166	3407	gp|L45171|	sialoglycoprotease [Haemophilus influenzae] >gp|U32735|HIU32735_5 sialoglycoprotease [Haemophilus influenzae] >gp|U00074|HIU00074_51 sialoglycoprotease [Haemophilus influenzae] >gp|U32844|HIU32844_5 sialoglycoprotease [Haemophilus influenzae] >pir|S|H6407	67	53
36	14	7522	9255	gp|M27221|	glutamyl-tRNA synthetase [Rhizobium meliloti] >pir|S|SYRZET glutamate--tRNA ligase (EC 6.1.1.17) - Rhizobium meliloti	67	48
45	6	6922	7989	gp|V01316|	tRNA synthetase [Saccharomyces cerevisiae] >gp|J01339|YSCMES1_1 Yeast (S.cerevisiae) methionyl-tRNA synthetase (mes1) gene, complete cds. [Saccharomyces cerevisiae] >gp|J01339|YSCMES1_1 S.cerevisiae methionyl-tRNA synthetase (mes1) gene, complete cds. [Sa	67	48
47	18	10834	10559	gp|X74121|	flagellar biosynthetic protein [Bacillus subtilis] >pir|S|S34714 flagellar protein flhB - Bacillus subtilis	67	48
51	4	4609	5790	gp|X94224|	DNA gyrase subunit A [Mycobacterium smegmatis] >gp|X94224|MSGYRBA_4 DNA gyrase subunit A [Mycobacterium smegmatis] >gp|X87117|MAPGYRA1_1 DNA gyrase [Mycobacterium abscessus] {SUB 75-114}	67	43
60	5	3353	2514	gp|J04836|	M.barkeri ATPase alpha and beta subunit (atpA and atpB) genes, complete cds. [Methanosarcina barkeri] >pir|S|A34283 H+-transporting ATP synthase (EC 3.6.1.34) alpha chain - Methanosarcina barkeri	67	51
62	1	572	3	gp|X73141|	hemolysin [Serpulina hyodysenteriae]	67	41

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

76	2	2197	677	gp|L27797|	DNA topoisomerase I [Bacillus subtilis]	67	51
91	1	1	426	gp|M88581|	transfer RNA-Leu synthetase [Bacillus subtilis] >pir|S|A41882 leucine--tRNA ligase (EC 6.1.1.4) - Bacillus subtilis	67	49
101	9	3437	3616	gp|L20784|	acetyl-CoA carboxylase [Cyclotella cryptica] >pir|S|A48757 acetyl-CoA carboxylase (EC 6.4.1.2) - Cyclotella cryptica	67	37
103	4	3167	4150	gp|U17902|	NtrC/NifA-like protein regulator [Escherichia coli]	67	47
113	5	2860	3357	gp|U21192|	SecA [Streptomyces lividans]	67	53
121	14	7227	7811	gp|X81642|	orf gene product [Wolinella succinogenes] >pir|S|S50154 hypothetical protein - Wolinella succinogenes	67	47
135	7	3256	3681	gp|L45521|	D-alanine permease [Haemophilus influenzae] >gp|U32770|HIU32770_3 D-alanine permease [Haemophilus influenzae] >gp|U00078|HIU00078_28 D-alanine permease [Haemophilus influenzae] >gp|U32716|HIU32716_5 alanine permease [Haemophilus influenzae] >pir|S|H64099	67	43
136	6	4442	3414	gp|M63176|	helicase [Staphylococcus aureus] >pir|S|S27667 DNA helicase pcrA - Staphylococcus aureus >pir|S|S39923 pcrA protein - Staphylococcus aureus	67	43
139	1	345	4	gp|Z28201|	S.cerevisiae chromosome XI reading frame ORF YKL202w. [Saccharomyces cerevisiae] >pir|S|S38038 hypothetical protein YKL201c - yeast (Saccharomyces cerevisiae)	67	41
140	1	1	345	gp|L45893|	periplasmic serine protease Do and heat shock protein [Haemophilus influenzae] >gp|U32805|HIU32805_12 periplasmic serine protease Do and heat shock protein [Haemophilus influenzae] >gp|U00082|HIU00082_28 periplasmic serine protease Do and heat shock prote	67	44
141	1	2	2029	gp|M58656|	pyruvate,orthophosphate dikinase [Zea mays] >gp|S46965|S46964S2_1 orthophosphate dikinase [Chloroplast Zea sp.] {SUB 1-154}	67	56
158	4	1690	1217	gp|U42580|	A7L gene product [Paramecium bursaria Chlorella virus 1]	67	39
165	4	922	1485	gp|L08626|	abc gene product [Escherichia coli] >gp|D83536|ECOTSF_24 ATP-binding protein [Escherichia coli]	67	51
213	1	257	48	gp|L38424|	polyA polymerase [Bacillus subtilis] >gp|L38424|BACJOJC_7	67	54

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					polyA polymerase [Bacillus subtilis]		
2	12	9394	9984	gp|D26185|	unknown [Bacillus subtilis]	66	39
4	11	7319	7885	gp|L47971|	methionine aminopeptidase [Bacillus subtilis] >gp|D00619|BACSECY_5 B.subtilis secY gene. [Bacillus subtilis] >pir|S|JS0493 methionyl aminopeptidase (EC 3.4.11.18) - Bacillus subtilis	66	42
8	12	6872	6351	gp|M90060|	Streptococcus faecalis H+ ATPase a (atpB),b (atpF),c (atpE),alpha (atpA), beta (atpD),gamma (atpG),delta (atpH),and epsilon (atpC) subunits, complete cds. [Streptococcus faecalis]	66	43
8	34	19445	20911	gp|D14162|	SecY protein [Corynebacterium glutamicum]	66	44
8	44	26260	26472	gp|X67646|	heat-shock protein [Borrelia burgdorferi] >gp|M96847|BORGRPEPLS_2 dnaK homologue gene product [Borrelia burgdorferi] >gp|M97912|BORHSP70A_1 70 kDa heat shock protein [Borrelia burgdorferi] >gp|S42385|S42385_1 HSP70 homolog [Borrelia burgdorferi, CA12 isol	66	43
16	4	2774	1785	gp|Z15056|	murE gene product [Bacillus subtilis] >pir|S|B47691 UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase (EC 6.3.2.13) - Bacillus subtilis	66	46
16	30	18350	17628	gp|L45639|	hypothetical protein (GB:L10328_69) [Haemophilus influenzae] >gp|U32781|HIU32781_4 hypothetical protein (GB:L10328_69) [Haemophilus influenzae] >gp|U00079|HIU00079_61 hypothetical protein (GB:L10328_69) [Haemophilus influenzae] >gp|U32727|HIU32727_5 inner	66	48
19	1	586	2	gp|L18927|	DNA polymerase III epsilon subunit [Buchnera aphidicola]	66	49
20	1	422	1105	gp|U00039|	E. coli chromosomal region from 76.0 to 81.5 minutes. [Escherichia coli] >pir|S|S47705 hypothetical protein f648 - Escherichia coli	66	48
21	1	676	1398	gp|J05534|	Escherichia coli ATP-dependent clp protease proteolytic component (clpP) gene, complete cds. [Escherichia coli] >pir|S|B36575 ATP-dependent Clp proteinase (EC 3.4.21.-) chain P precursor - Escherichia coli	66	44
21	9	7147	8034	gp|D64006|	hypothetical protein [Synechocystis sp.]	66	48

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

27	4	1615	2550	gp|Z49782|	peptide chain release factor 1 [Bacillus subtilis] >pir|S|S55437 peptide chain release factor 1 - Bacillus subtilis	66	44
28	4	3054	1852	gp|L40822|	phophoglucoisomerase-like protein [Chlamydia trachomatis]	66	50
28	7	3775	3392	gp|L23147|	acetate kinase [Methanosarcina thermophila] >pir|S|B49338 acetate kinase (EC 2.7.2.1) - Methanosarcina thermophila	66	40
35	32	18159	17989	gp|U10577|	bone sialoprotein II [Gallus gallus]	66	53
35	34	18290	18144	gp|D43920|	DNA (cytosine-5-) -methyltransferase [Gallus gallus] >gp|D43920|CHKMETASE_1 DNA (cytosine-5-) - methyltransferase [Gallus gallus]	66	40
39	3	950	777	gp|L47164|	dnrQ gene product [Streptomyces peucetius] >gp|L47164|STMDNRQ_1 dnrP gene product [Streptomyces peucetius]	66	46
47	12	6842	5379	gp|U18744|	MgtE [Bacillus firmus]	66	39
48	6	3736	3029	gp|L27492|	triosephosphate isomerase [Thermotoga maritima]	66	50
52	6	6983	6273	gp|X73141|	hemolysin [Serpulina hyodysenteriae]	66	40
53	9	5378	3882	gp|M29495|	unknown protein [Synechococcus sp.] >pir|S|Q3YCRQ hypothetical protein (recA 3' region) - Synechococcus sp. (PCC 7002) (fragment)	66	35
57	3	1788	2258	gp|J03218|	S.acidocaldarius membrane-associated ATPase alpha subunit gene, complete cds. [Sulfolobus acidocaldarius] >pir|S|A28652 H+-transporting ATP synthase (EC 3.6.1.34) alpha chain, membrane-associated - Sulfolobus acidocaldarius	66	42
57	5	3732	4421	gp|U47274|	A1AO H+ ATPase, subunit D [Methanosarcina mazeii]	66	39
58	3	1805	1167	gp|U11045|	ORF-C gene product [Buchnera aphidicola]	66	42
61	13	8053	7388	gp|U45426|	heat shock protein HTPG [Borrelia burgdorferi] >gp|L32145|BORHTPG_1 C62.5 heat shock protein [Borrelia burgdorferi] {SUB 497-575}	66	37
62	4	1966	2160	gp|U19831|	SpZ12-1 [Strongylocentrotus purpuratus]	66	53
63	1	2	979	gp|U27343|	MutL [Bacillus subtilis]	66	42
67	2	2605	1454	pir|S|A48018	mucin, MG2=low molecular weight salivary glycoprotein - human >pir|S|S29115 mucin MG2b-T2 - human {SUB 143-168}	66	47
76	1	740	3	gp|L27797|	DNA topoisomerase I [Bacillus subtilis]	66	43

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

121	5	1550	1341	gp|U20669|	ribosomal protein S21 [Myxococcus xanthus]	66	45
121	8	3037	3657	gp|L09703|	penicillin-binding protein [Bacillus subtilis] >gp|Z68230|BSYLLSPO_5 high molecular weight penicillin binding protein [Bacillus subtilis] >pir|S|C53292 penicillin-binding protein 2B - Bacillus subtilis >gp|Z25865|BSSPOVD_1 Pbp2B [Bacillus subtilis] {SUB 6	66	37
127	1	636	4	gp|L44828|	hypothetical protein (SP:P30143) [Haemophilus influenzae] >gp|U32703|HIU32703_7 hypothetical protein (SP:P30143) [Haemophilus influenzae] >gp|U00070|HIU00070_87 hypothetical protein (SP:P30143) [Haemophilus influenzae] >gp|U32812|HIU32812_7 amino acid per	66	57
129	2	1410	2207	gp|U29399|	major outer sheath protein [Treponema denticola]	66	36
131	1	471	4	gp|U04953|	isoleucyl-tRNA synthetase [Homo sapiens]	66	49
141	3	2103	2453	gp|X59528|	L20 gene product [Bacillus subtilis] >pir|S|S18439 Ribosomal protein L21 - Bacillus subtilis	66	41
158	3	702	463	gp|U19861|	mab-21 gene product [Caenorhabditis elegans]	66	66
165	1	190	2	gp|U14003|	Escherichia coli K-12 chromosomal region from 92.8 to 00.1 minutes. [Escherichia coli] >pir|S|S56374 hypothetical protein f342 - Escherichia coli	66	52
266	1	86	289	pir|S|RRWV EV	RNA-directed RNA polymerase (EC 2.7.7.48) - equine arteritis virus >gp|X53459|TOEAV_1 Equine arteritis virus (EAV) RNA genome. [Equine arteritis virus] {SUB 1-1727}	66	44
2	15	13965	13210	gp|D26185|	transcription-repair coupling factor [Bacillus subtilis]	65	41
5	1	74	457	gp|D83536|	unknown [Escherichia coli]	65	52
8	26	13905	14519	gp|U43929|	S3 [Bacillus subtilis]	65	47
8	30	17102	17455	gp|X79551|	ribosomal S8 protein [Thermus aquaticus thermophilus] >pir|S|A53870 ribosomal protein S8 - Thermus aquaticus >pir|S|S51059 ribosomal protein S8 - Thermus aquaticus {SUB 1-28}	65	59
8	31	17453	18004	pir|S|R5BS0 F	ribosomal protein L6 - Bacillus stearothermophilus	65	48
16	36	22233	21427	pir|S|A4166	hypothetical protein 1 - Enterococcus faecalis plasmid pCF10	65	40

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

				2	(fragment)		
19	3	2367	2894	gp|U18937|	histidyl-tRNA synthetase homologue [Homo sapiens] >gp|U18936|HSU18936_1 histidyl-tRNA synthetase homologue [Homo sapiens] {SUB 1-36}	65	42
19	16	13750	13091	pir|S|JC1317	protein-methionine-S-oxide reductase (EC 1.8.4.6) - Escherichia coli	65	53
21	18	11207	10506	gp|U18997|	Escherichia coli K-12 chromosomal region from 67.4 to 76.0 minutes. [Escherichia coli]	65	44
29	26	12129	12341	gp|S79915|	Hls [Drosophila]	65	38
33	2	372	677	gp|M69036|	protein H [Alcaligenes eutrophus] >pir|S|A38120 phbH protein - Alcaligenes eutrophus	65	42
35	14	8692	9093	gp|X87899|	thioredoxin /thioredoxin reductase hybrid protein [Mycobacterium leprae] >gp|L39923|MSGDNAB_14 thioredoxin reductase/thioredoxin [Mycobacterium leprae]	65	42
35	33	18117	18719	gp|L46071|	hypothetical protein (SP:P26242) [Haemophilus influenzae] >gp|U32822|HIU32822_12 hypothetical protein (SP:P26242) [Haemophilus influenzae] >gp|U00084|HIU00084_34 hypothetical protein (SP:P26242) [Haemophilus influenzae] >gp|U32768|HIU32768_15 transketolas	65	41
36	18	11816	13033	gp|M20793|	S.typhimurium D-alanine:D-alanine ligase (ddlA) gene, complete cds. [Salmonella typhimurium] >pir|S|CEEBDT D-alanine--D-alanine ligase (EC 6.3.2.4) A - Salmonella typhimurium	65	44
42	8	3884	4252	gp|X02499|	Rhodospirillum rubrum atp operon. [Rhodospirillum rubrum] >pir|S|S08579 hypothetical protein 2 - Rhodospirillum rubrum >gp|X02499|RRATP_4 Rhodospirillum rubrum atp operon. [Rhodospirillum rubrum] {SUB 592-811}	65	48
44	6	3647	2583	gp|X95669|	thdF gene product [Borrelia burgdorferi] >gp|Z12160|BBGIDAG_1 thdF gene product [Borrelia burgdorferi] {SUB 429-463}	65	50
47	4	2157	1501	gp|L47709|	ypiA gene product [Bacillus subtilis]	65	50
47	20	12498	12091	gp|L44847|	tRNA (guanine-N1)-methyltransferase [Haemophilus influenzae] >gp|U32705|HIU32705_6 tRNA (guanine-N1)-methyltransferase	65	47

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					[Haemophilus influenzae] >gp|U00071|HIU00071_17 tRNA (guanine-N1)-methyltransferase [Haemophilus influenzae] >gp|U32814|HIU32814_6 tRNA		
48	9	6125	7117	gp|U18532|	Bex [Bacillus subtilis]	65	40
70	3	908	2131	gp|X73124|	ipa-65d gene product [Bacillus subtilis] >pir|S|S39720 hypothetical protein - Bacillus subtilis	65	44
72	1	314	18	gp|U09005|	HflC [Vibrio parahaemolyticus]	65	34
79	5	2430	3350	gp|U43739|	FliF [Borrelia burgdorferi] >gp|L76303|BORFTSA_10 flagellar basal body rod protein [Borrelia burgdorferi]	65	42
81	9	5977	7617	gp|M91593|	Mycoplasma mycoides SRPM54 gene, complete cds. [Mycoplasma mycoides] >pir|S|S35480 hypothetical protein 1 - Mycoplasma mycoides (SGC3) >pir|S|S27590 hypothetical protein 1 - Mycoplasma mycoides (SGC3) (fragment) {SUB 2-422}	65	40
83	13	4268	4939	pir|S|PC2307	X-Pro aminopeptidase (EC 3.4.11.9) L13K - guinea pig (fragment) >pir|S|PC2310 X-Pro aminopeptidase (EC 3.4.11.9) K13K - guinea pig (fragment) {SUB 1-26}	65	51
85	7	3480	2587	gp|M64519|	transport protein [Escherichia coli] >pir|S|C40840 spermidine/putrescine transmembrane protein C - Escherichia coli	65	35
88	8	3582	3902	gp|X77636|	putative amino acid binding subunit [Bacillus subtilis] >pir|S|S52381 probable amino acid binding protein - Bacillus subtilis	65	37
89	5	2189	2815	gp|M30198|	recQ gene product [Escherichia coli]	65	46
98	2	1233	760	gp|X77925|	dUTPase [Candida albicans] >pir|S|S42871 dUTP pyrophosphatase (EC 3.6.1.23) - yeast (Candida albicans)	65	46
101	1	440	3	gp|U01322|	ribonucleotide reductase small subunit [Plasmodium falciparum] >pir|S|B49412 ribonucleoside-diphosphate reductase (EC 1.17.4.1) small subunit (EC 1.17.4.1) - Plasmodium falciparum	65	40
104	15	5605	6159	gp|M13462|	cheW gene product [Escherichia coli] >pir|S|QRECCW chemotaxis protein cheW - Escherichia coli	65	43
115	9	4532	4167	gp|J05478|	alkyl hydroperoxide reductase [Salmonella typhimurium]	65	47
121	1	774	208	gp|X79087|	methionyl-tRNA formyltransferase [Thermus aquaticus thermophilus] >pir|S|B55228 fmt protein homolog - Thermus	65	43

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					aquaticus		
121	18	8556	8110	gp|D26185|	cell division protein [Bacillus subtilis]	65	46
125	2	1975	1763	gp|M76442|	crystal protein [Bacillus thuringiensis] >pir|S|A41969 crystal protein, 40K - Bacillus thuringiensis >gp|M90843|BACCRY_1 crystal protein [Bacillus thuringiensis] {SUB 1-39}	65	35
144	4	1050	649	gp|X69601|	p93 [Borrelia burgdorferi]	65	37
150	6	4798	5175	gp|L47709|	poly(A) polymerase [Bacillus subtilis]	65	50
150	8	7329	6343	gp|D26185|	stage 0 sporultion [Bacillus subtilis] >gp|X62539|BSORIGS_11 spoOJ93 gene product [Bacillus subtilis] >pir|S|A38536 spo0J93 protein - Bacillus subtilis	65	47
151	1	1231	2	gp|M30297|	B.subtilis recombination and sporulation protein (recN, spoIVB) genes , complete cds, arginine hydroximate resistance (ahrC) gene, 3' end. [Bacillus subtilis] >pir|S|B35128 recN homolog - Bacillus subtilis	65	46
1	7	2258	3169	gp|D83536|	proline-tRNA ligase [Escherichia coli]	64	47
1	10	3063	3620	gp|L25105|	aminoacyl-tRNA synthetase [Chlamydia trachomatis]	64	41
2	5	2357	4033	gp|L15202|	Bacillus subtilis comE operon encoding ORF1, ORF2, ORF3 and Reverse-ORF genes, complete cds. [Bacillus subtilis] >pir|S|S39865 ComE ORF3 - Bacillus subtilis	64	31
4	7	4689	5108	gp|U39483|	EMG2 [Escherichia coli]	64	45
5	3	1285	2172	gp|D64000|	hypothetical protein [Synechocystis sp.]	64	46
8	40	23622	24191	gp|D28550|	heat shock protein GrpE homolog [Synechococcus sp.] >pir|S|PC2235 GrpE protein - Synechococcus sp. (PCC 7942) (fragment)	64	43
8	42	23945	24448	gp|M84964|	heat shock protein [Bacillus subtilis] >gp|X51477|BSGRPE_1 Bacillus subtilis DNA for grpE gene and dnaK gene (partial). [Bacillus subtilis] >pir|S|S08418 heat shock protein grpE - Bacillus subtilis	64	38
11	19	8921	9778	gp|U29399|	major outer sheath protein [Treponema denticola]	64	38
16	20	13266	12769	gp|M85240|	flagellar protein [Escherichia coli]	64	33
16	24	15274	14606	gp|Z50098|	pentose-5-phosphate-3-epimerase [Solanum tuberosum]	64	44
37	10	4141	4788	gp|U18997|	Escherichia coli K-12 chromosomal region from 67.4 to 76.0	64	40

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					minutes. [Escherichia coli]		
41	14	5354	5845	gp|D64116|	Srb [Bacillus subtilis] >gp|D49781|BACSRBA_2 Srb [Bacillus subtilis] >pir|S|JC4093 signal recognition particle receptor alpha chain homolog - Bacillus subtilis	64	46
47	11	5268	5486	pir|S|A40701	tenascin-X precursor - human >gp|X71937|HSXBXVIII_1 fibrinogen [Homo sapiens] {SUB 3356-3566} >pir|S|C42175 tenascin homolog 3.9kF3-1 - human (fragment) {SUB 1849-1936}	64	42
54	27	8658	8260	gp|M60917|	purine nucleoside phosphorylase [Escherichia coli] >gp|U14003|ECOUW93_295 purine-nucleoside phosphorylase [Escherichia coli] >pir|S|A27854 purine-nucleoside phosphorylase (EC 2.4.2.1) - Escherichia coli	64	51
58	1	575	3	gp|X84019|	orf3 gene product [Zymomonas mobilis] >gp|X84019|ZMDNAGRP_3 orf3 gene product [Zymomonas mobilis]	64	36
62	10	4019	4354	gp|L45980|	spermidine/putrescine transport ATP-binding protein [Haemophilus influenzae] >gp|U32813|HIU32813_12 spermidine/putrescine transport ATP-binding protein [Haemophilus influenzae] >gp|U00083|HIU00083_23 spermidine/putrescine transport ATP-binding protein [Ha	64	49
64	21	9281	9610	gp|Z54238|	T28C6.1 [Caenorhabditis elegans]	64	52
70	23	15605	15985	gp|D26185|	cysteinyl-tRNA synthetase [Bacillus subtilis] >gp|L14580|BACGLUSYN_6 cysteinyl-tRNA synthetase [Bacillus subtilis] >gp|X73989|BSCTS_1 cysteine--tRNA ligase [Bacillus subtilis] >pir|S|C53402 cysteine--tRNA ligase (EC 6.1.1.16) - Bacillus subtilis	64	50
88	4	2017	2583	gp|U18539|	FliY [Escherichia coli]	64	33
100	42	18967	20169	gp|X81642|	orf gene product [Wolinella succinogenes] >pir|S|S50154 hypothetical protein - Wolinella succinogenes	64	48
112	15	10532	9804	gp|U30821|	putative protein of 244 amino acids [Cyanelle Cyanophora paradoxa]	64	42
115	7	2550	3152	gp|X91655|	lepA gene product [Bacillus subtilis] >gp|D17650|BACGPR_4	64	45

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					ORF80 protein [Bacillus subtilis] {SUB 1-327}		
121	27	15390	17069	gp|L44876|	ATP-dependent RNA helicase [Haemophilus influenzae] >gp|U32709|HIU32709_6 ATP-dependent RNA helicase [Haemophilus influenzae] >gp|U00071|HIU00071_47 ATP-dependent RNA helicase [Haemophilus influenzae] >gp|U32818|HIU32818_6 RNA (?) helicase [Haemophilus in	64	47
123	1	44	439	gp|U29399|	major outer sheath protein [Treponema denticola]	64	51
129	1	1208	3	gp|U29399|	major outer sheath protein [Treponema denticola]	64	38
170	1	874	92	gp|M54884|	xprB gene product [Escherichia coli] >gp|U28375|ECU28375_44 site-specific integrase/recombinase, with xerC [Escherichia coli] >pir|S|A39202 recombinase XerD - Escherichia coli	64	42
552	1	3	407	gp|M27869|	B.subtilis ahrC gene, encoding an arginine repressor/activator protein . [Bacillus subtilis]	64	39
636	2	265	642	gp|M26414|	RNA polymerase alpha-core-subunit [Bacillus subtilis] >gp|L47971|BACRPLP_21 RNA polymerase alpha-core-subunit [Bacillus subtilis] >pir|S|E32307 DNA-directed RNA polymerase (EC 2.7.7.6) alpha chain - Bacillus subtilis >gp|M13957|BACRPOA_3 B.subtilis DNA se	64	40
685	1	2	286	gp|X62681|	limb deformity protein [Gallus gallus] >pir|S|S38780 limb deformity protein - chicken	64	64
693	2	383	168	gp|U41047|	Genesis [Mus musculus] >gp|L13202|RATHFH2_1 HNF-3/fork-head homolog-2 [Rattus norvegicus] {SUB 129-229}	64	58
2	9	5368	6384	gp|L44821|	hypothetical protein (SP:P33643) [Haemophilus influenzae] >gp|U32702|HIU32702_15 hypothetical protein (SP:P33643) [Haemophilus influenzae] >gp|U00070|HIU00070_80 hypothetical protein (SP:P33643) [Haemophilus influenzae] >gp|U32811|HIU32811_16 pseudoU synt	63	38
8	25	13565	13933	gp|Z21677|	ribosomal protein L22 [Thermotoga maritima] >pir|S|S40193 ribosomal protein L22 - Thermotoga maritima	63	45
8	47	27536	27841	gp|D64006|	hypothetical protein [Synechocystis sp.]	63	36
16	29	17481	16768	gp|D26185|	unknown [Bacillus subtilis] >gp|X62539|BSORIGS_4 B.subtilis genes rpmH, rnpA, 50kd, gidA and gidB. [Bacillus subtilis]	63	39

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					>gp|Z14225|BSRNPASPO_3 Jag [Bacillus subtilis] >pir|S|S18074 jag protein - Bacillus subtilis		
21	12	8978	9610	gp|J03497|	valyl tRNA synthetase [Escherichia coli]	63	44
22	17	11185	11997	gp|U29580|	mazG gene product [Escherichia coli]	63	45
22	19	13409	13194	gp|L46319|	endonuclease III [Haemophilus influenzae] >gp|U32842|HIU32842_1 endonuclease III [Haemophilus influenzae] >gp|U00086|HIU00086_51 endonuclease III [Haemophilus influenzae] >gp|U32788|HIU32788_7 endonuclease III [Haemophilus influenzae] >pir|S|G64136 endonu	63	49
29	22	10192	9176	gp|U29399|	major outer sheath protein [Treponema denticola]	63	43
39	9	4947	3727	gp|U26401|	galactokinase [Homo sapiens] >gp|U26401|HSU26401_1 galactokinase [Homo sapiens]	63	48
42	9	4123	4872	gp|L44983|	primosomal protein replication factor [Haemophilus influenzae] >gp|U32718|HIU32718_10 primosomal protein replication factor [Haemophilus influenzae] >gp|U00072|HIU00072_68 primosomal protein replication factor [Haemophilus influenzae] >gp|U32827|HIU32827_	63	45
47	26	15583	14249	gp|Z15056|	murD gene product [Bacillus subtilis] >pir|S|D47691 UDP-N-acetylmuramoylalanine--D-glutamate ligase (EC 6.3.2.9) - Bacillus subtilis	63	42
49	8	6603	5830	gp|L45045|	hypothetical protein (GB:D26185_130) [Haemophilus influenzae] >gp|U32724|HIU32724_5 hypothetical protein (GB:D26185_130) [Haemophilus influenzae] >gp|U00073|HIU00073_26 hypothetical protein (GB:D26185_130) [Haemophilus influenzae] >gp|U32832|HIU32832_13 A	63	40
54	18	6674	6345	pir|S|B6122 8	collagen alpha 1(IV) chain - rabbit (fragment)	63	40
56	8	3790	4077	gp|D45163|	embryonic muscle myosin heavy chain [Halocynthia roretzi]	63	27
58	4	2308	1829	gp|L46186|	H. influenzae predicted coding region HI1555 [Haemophilus influenzae] >gp|U32830|HIU32830_11 H. influenzae predicted coding region HI1555 [Haemophilus influenzae] >gp|U00085|HIU00085_26 H. influenzae predicted coding region	63	33

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					HI1555 [Haemophilus influenzae]		
61	12	6015	6527	gpIU18997I	glpG gene product [Escherichia coli]	63	32
69	13	7256	7561	gpIX69377I	glycine-rich RNA-binding protein [Arabidopsis thaliana] >pirIS1S31443 glycine-rich RNA-binding protein (clone A81) - Arabidopsis thaliana (fragment) >gpIZ18189IATTS0693_1 GLYCINE-RICH RNA-BINDING PROTEIN [Arabidopsis thaliana] {SUB 1-66}	63	41
77	2	904	1065	pirIS1TNLJS2	trans-activating transcriptional regulatory protein - simian immunodeficiency virus SIVagm (type 3, isolate STLV-3agm)	63	47
81	19	16307	17188	gpIM87049I	uridine phosphorylase [Escherichia coli] >gpIX15689IECUDP_1 E. coli udp gene for uridine phosphorylase (EC 2.4.2.3). [Escherichia coli] >pirIS1S05491 uridine phosphorylase (EC 2.4.2.3) - Escherichia coli	63	48
88	1	1	966	gpIL45439I	ribosomal protein S4 [Haemophilus influenzae] >gpIU32762IHIU32762_15 ribosomal protein S4 [Haemophilus influenzae] >gpIU00077IHIU00077_61 ribosomal protein S4 [Haemophilus influenzae] >gpIU32708IHIU32708_16 ribosomal protein S4 [Haemophilus influenzae] >p	63	40
111	2	1345	1902	gpIU12513I	PPi-dependent phosphofructo-1-kinase [Entamoeba histolytica] >gpIU12513IEHU12513_1 PPi-dependent phosphofructo-1-kinase [Entamoeba histolytica] >pirIS1S52082 PPi-dependent phosphofructo-1-kinase - Entamoeba histolytica	63	52
121	4	1277	1113	gpIL45263I	polypeptide deformylase (formylmethionine deformylase) [Haemophilus influenzae] >gpIU32745IHIU32745_1 polypeptide deformylase (formylmethionine deformylase) [Haemophilus influenzae] >gpIU00075IHIU00075_42 polypeptide deformylase (formylmethionine deformyl	63	44
167	3	1293	1490	gpID26562I	'ribosome releasing factor' [Escherichia coli] >gpIJ05113IECORRFX_1 E.coli ribosome releasing factor gene, complete cds. [Escherichia coli] >gpID13334IECOSMBA_3 ribosome releasing factor (FRR) [Escherichia coli] >gpID83536IECOTSF_3 ribosome releasing fact	63	42
646	2	417	674	gpID37850I	core, env,and part of E2/NS1 [Hepatitis C virus]	63	47

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					>gp|D37850|HPCCORE13_1 core, env and part of E2/NS1 [Hepatitis C virus]		
5	5	2360	3250	gp|D83536|	unknown [Escherichia coli]	62	45
10	18	9417	9190	gp|U26536|	glycerophosphoryl diester phosphodiesterase [Escherichia coli]	62	40
29	6	2643	2431	gp|X94335|	YOR3174c gene product [Saccharomyces cerevisiae]	62	37
38	3	631	1548	gp|Z70306|	T25B9.9 [Caenorhabditis elegans]	62	47
42	1	999	679	gp|U34795|	GlnQ [Mycoplasma pneumoniae]	62	36
42	13	7604	7918	gp|L10348|	transcription factor [Thermus thermophilus] >pir|S|A45281 transcription factor NusG - Thermus thermophilus {SUB 61-244}	62	50
44	24	14444	13665	gp|U33064|	transketolase [Xanthobacter flavus] >gp|U33064|XFU33064_1 transketolase [Xanthobacter flavus]	62	48
45	13	10253	11362	gp|L44817|	hypothetical protein (SP:P33943) [Haemophilus influenzae] >gp|U32702|HIU32702_11 hypothetical protein (SP:P33943) [Haemophilus influenzae] >gp|U00070|HIU00070_76 hypothetical protein (SP:P33943) [Haemophilus influenzae] >gp|U32811|HIU32811_12 reductase su	62	40
48	19	12583	11879	gp|X59389|	galactose binding protein [Citrobacter freundii] >pir|S|S15554 galactose-binding protein - Citrobacter freundii	62	45
55	1	2	475	gp|U10397|	YHR155w gene product [Saccharomyces cerevisiae] >pir|S|S46754 hypothetical protein YHR155w - yeast (Saccharomyces cerevisiae)	62	31
62	3	1895	945	gp|D64004|	hypothetical protein [Synechocystis sp.]	62	42
64	4	1781	2050	pir|S|A39401	merozoite surface antigen 1 precursor - Plasmodium vivax	62	41
69	5	2474	1995	gp|Z21496|	histidine rich protein [Escherichia coli] >gp|L28082|ECOSLYD_1 slyD gene product [Escherichia coli] >gp|L13261|ECOSLYDX_2 slyD gene product [Escherichia coli] >gp|U18997|ECOUW67_273 histidine rich protein [Escherichia coli] >pir|S|A49987 metal-binding pro	62	43
69	12	6034	6690	gp|D63999|	hypothetical protein [Synechocystis sp.]	62	37
70	21	14901	15395	pir|S|F53402	cysteine--tRNA ligase (EC 6.1.1.16) - Bacillus stearothermophilus (fragment)	62	55

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

83	8	3198	3713	gp|Z54328|	unknown [Schizosaccharomyces pombe]	62	44
87	1	1678	2	gp|S57688|	EF-G [Thermotoga maritima]	62	44
89	13	11242	11454	gp|U17013|	Oct1 [Rattus norvegicus]	62	31
91	3	902	1405	gp|M88581|	transfer RNA-Leu synthetase [Bacillus subtilis] >pir|S|A41882 leucine--tRNA ligase (EC 6.1.1.4) - Bacillus subtilis	62	46
94	5	2441	2160	gp|D10483|	ORF [Escherichia coli] >pir|S|QQECFT hypothetical 38.8K protein (ftsI 5' region) - Escherichia coli >gp|X55034|EC2MIN_9 E. coli 2 minute region. [Escherichia coli] {SUB 34-346}	62	43
95	3	1626	277	gp|U29399|	major outer sheath protein [Treponema denticola]	62	43
102	5	3413	2340	gp|X02164|	E.coli ponA gene for penicillin-binding protein 1A (PBP 1A). [Escherichia coli] >pir|S|ZPECPA penicillin-binding protein 1A - Escherichia coli	62	40
104	19	7930	8310	gp|U36840|	Escherichia coli K-12 genome, approximately 57 minutes. [Escherichia coli]	62	46
110	1	16	429	gp|X64451|	gcpE gene product [Escherichia coli] >pir|S|S23058 gcpE protein - Escherichia coli	62	51
111	1	698	1003	gp|Z26653|	laminin M chain (merosin) [Homo sapiens]	62	33
115	1	803	231	gp|D26185|	unknown [Bacillus subtilis] >gp|L14580|BACGLUSYN_3 unknown [Bacillus subtilis] >pir|S|A53402 glutamate--tRNA ligase (EC 6.1.1.17) - Bacillus subtilis (fragment) {SUB 137-158}	62	47
116	7	2360	1719	gp|J05015|	Synthetic Clostridium MP flavodoxin gene, complete cds. [Artificial gene] >pir|S|FXCLEX flavodoxin - Clostridium sp.	62	47
116	9	3316	2573	gp|L09189|	Neisseria meningitidis dTDP-D-glucose 4,6-dehydratase (rfbB), glucose-1-phosphate thymidyl transferase (rfbA) and rfbC genes, complete cds and UPD-glucose-4-epimerase (galE) pseudogene. [Neisseria meningitidis]	62	39
141	5	3195	3944	gp|M24537|	GTP-binding protein [Bacillus subtilis] >pir|S|B32804 GTP-binding protein, spo0B 3'-region - Bacillus subtilis >gp|K02666|BACSPO0B2_2 Bacillus subtilis spo0B early sporulation gene, complete cds. [Bacillus subtilis] {SUB 1-65}	62	48
152	1	3	1964	gp|J03294|	uvrB gene product [Bacillus subtilis] >gp|J03294|BACAPKII_2 deoxyribodipyrimidine photolyase [Bacillus subtilis]	62	48

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					>pirlSlA37192 uvrB protein - Bacillus subtilis		
160	2	272	874	gplZ12160l	gidA gene product [Borrelia burgdorferi] >gplZ12160lBBGIDAG_1 division protein [Borrelia burgdorferi] {SUB 529-593} >gplX95669lBBTHDFGID_2 gidA gene product [Borrelia burgdorferi] {SUB 1-29} >gplX95668lBBGIDMOXR_1 gidA gene product [Borrelia burgdorferi]	62	34
194	1	378	4	gplU04953l	isoleucyl-tRNA synthetase [Homo sapiens]	62	44
574	1	439	20	gplD26185l	unknown [Bacillus subtilis] >gplU02604lBSU02604_3 ORFY [Bacillus subtilis] {SUB 1-260}	62	40
599	1	1	321	gplD26185l	unknown [Bacillus subtilis]	62	37
645	2	592	200	gplU29399l	major outer sheath protein [Treponema denticola]	62	37
4	12	7734	8114	gplL47971l	methionine aminopeptidase [Bacillus subtilis] >gplD00619lBACSECY_5 B.subtilis secY gene. [Bacillus subtilis] >pirlSlJS0493 methionyl aminopeptidase (EC 3.4.11.18) - Bacillus subtilis	61	47
8	22	11941	12378	gplZ21677l	ribosomal protein L23 [Thermotoga maritima] >pirlSlS40190 ribosomal protein L23 - Thermotoga maritima	61	41
13	12	7268	7047	gplZ46792l	C09G5.8 [Caenorhabditis elegans]	61	38
30	17	12484	13842	gplM22857l	rodA gene product [Escherichia coli] >pirlSlBVECRD rod shape-determining protein mrdB - Escherichia coli	61	39
35	30	17342	18001	pirlSlS15618	E4 protein - human papillomavirus type 2a	61	47
35	41	24978	24346	gplU51032l	D9651.10 gene product [Saccharomyces cerevisiae]	61	38
44	14	8614	7922	gplL42115l	insulin-activated amino acid transporter [Mus musculus] >pirlSlJC4149 adipocyte amino acid transporter - mouse	61	38
49	12	8668	8234	gplD26185l	unknown [Bacillus subtilis]	61	46
57	8	7197	6619	gplD64001l	hypothetical protein [Synechocystis sp.]	61	37
63	27	16398	14845	gplX73124l	ipa-68d gene product [Bacillus subtilis] >pirlSlS39723 hypothetical protein - Bacillus subtilis	61	44
65	2	396	214	gplL14745l	C. elegans cosmid C02F5. [Caenorhabditis elegans] >pirlSlS44608 C02F5.6 protein - Caenorhabditis elegans	61	44
79	1	2	895	gplX78478l	sss gene product [Pseudomonas aeruginosa] >pirlSlS43156 sss protein - Pseudomonas aeruginosa	61	49

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

94	2	1580	750	gp|X62437|	UDP-N-acetylmuramoylalanyl-D-glutamyl-2, 6-diaminopimelate--D-alanyl-D-alanine ligase [Synechocystis sp.] >pir|S|S49610 UDP-N-acetylmuramoylalanyl-D-glutamyl-2, 6-diaminopimelate--D-alanyl-D-alanine ligase (EC 6.3.2.15) - Synechocystis sp. (PCC 6803)	61	43
102	1	934	101	gp|L45849|	single-stranded-DNA-specific exonuclease [Haemophilus influenzae] >gp|U32801|HIU32801_1 single-stranded-DNA-specific exonuclease [Haemophilus influenzae] >gp|U00081|HIU00081_78 single-stranded-DNA-specific exonuclease [Haemophilus influenzae] >gp|U32746|H	61	43
103	2	693	1931	gp|D26185|	similar to B. subtilis DnaH [Bacillus subtilis]	61	38
104	12	4994	5404	gp|U28377|	Escherichia coli K-12 genome; approximately 65 to 68 minutes. [Escherichia coli]	61	48
116	13	5011	4265	gp|U32847|	ribosomal protein S1 homolog, RNA-binding protein [Haemophilus influenzae]	61	39
117	5	3484	2504	gp|M34066|	trigger factor [Escherichia coli] >pir|S|A36129 trigger factor protein - Escherichia coli >gp|J05534|ECOCLPPA_1 Escherichia coli ATP-dependent clp protease proteolytic component (clpP) gene, complete cds. [Escherichia coli] {SUB 390-432}	61	29
121	16	8072	7830	gp|Z47047|	unknown [Saccharomyces cerevisiae] >gp|Z46902|SC8224_5 unknown [Saccharomyces cerevisiae] {SUB 87-316}	61	19
128	1	1415	135	gp|L26051|	UDP-N-acetylglucosamine 1-carboxyvinyl transferase [Acinetobacter calcoaceticus]	61	39
142	5	7013	8107	gp|D26185|	DNA polymerase III subunit [Bacillus subtilis] >gp|X17014|BSRECM_1 dnaZX gene product [Bacillus subtilis] >pir|S|S13786 DNA-directed DNA polymerase (EC 2.7.7.7) III chain dnaX - Bacillus subtilis >gp|X06803|BSDNAZX_1 Bacillus subtilis DNA for dnaZX-like O	61	43
150	7	5018	6346	gp|L47709|	poly(A) polymerase [Bacillus subtilis]	61	46
324	1	1	192	gp|L45521|	D-alanine permease [Haemophilus influenzae] >gp|U32770|HIU32770_3 D-alanine permease [Haemophilus influenzae] >gp|U00078|HIU00078_28 D-alanine permease [Haemophilus influenzae] >gp|U32716|HIU32716_5 alanine	61	42

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					permease [Haemophilus influenzae] >pir|S|H64099		
328	2	508	296	gp|X66504|	ExeL gene product [Aeromonas hydrophila] >pir|S|E49905 protein secretion operon exe protein L - Aeromonas hydrophila	61	38
681	2	239	466	gp|X16278|	Thermus thermophilus fus gene for elongation factor G (EF-G). [Thermus aquaticus thermophilus] >pir|S|EFTWG translation elongation factor G - Thermus aquaticus	61	33
8	37	22115	22876	gp|M26414|	RNA polymerase alpha-core-subunit [Bacillus subtilis] >gp|L47971|BACRPLP_21 RNA polymerase alpha-core-subunit [Bacillus subtilis] >pir|S|E32307 DNA-directed RNA polymerase (EC 2.7.7.6) alpha chain - Bacillus subtilis >gp|M13957|BACRPOA_3 B.subtilis DNA se	60	38
16	41	24204	23581	gp|X82071|	orf3 gene product [Pseudomonas aeruginosa] >pir|S|S49376 hypothetical protein 3 - Pseudomonas aeruginosa	60	26
16	48	28011	28598	pir|S|S19739	integral membrane protein - Rhodobacter capsulatus	60	35
21	10	7764	7165	gp|L44744|	H. influenzae predicted coding region HI0100 [Haemophilus influenzae] >gp|U32695|HIU32695_7 H. influenzae predicted coding region HI0100 [Haemophilus influenzae] >gp|U00070|HIU00070_7 H. influenzae predicted coding region HI0100 [Haemophilus influenzae] >	60	34
26	24	13879	14238	gp|M59444|	mglA gene product [Escherichia coli] >pir|S|B37277 50K membrane-associated protein mglA - Escherichia coli	60	44
35	7	4135	4671	gp|D64116|	ORF3 [Bacillus subtilis]	60	41
36	13	8256	7405	gp|D50417|	ATEC3 [Mus musculus]	60	39
41	3	915	1574	gp|M62785|	multiphosphoryl transfer protein [Rhodobacter capsulatus] >gp|X53150|RCFRUOP_1 multiphosphoryl transfer protein [Rhodobacter capsulatus] >pir|S|S10639 fruB protein - Rhodobacter capsulatus	60	45
44	7	4075	3578	gp|X95669|	thdF gene product [Borrelia burgdorferi] >gp|Z12160|BBGIDAG_1 thdF gene product [Borrelia burgdorferi] (SUB 429-463)	60	41
45	5	6624	5203	gp|X06480|	E.coli dacC gene for penicillin-binding protein 6. [Escherichia coli] >pir|S|B28536 penicillin-binding protein 6 precursor - Escherichia	60	49

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					coli		
46	2	1235	858	gplU45285l	specific 116-kDa vacuolar proton pump subunit [Homo sapiens]	60	40
47	22	12967	12641	gplU39731l	tRNA (guanine-N1)-methyltransferase [Mycoplasma genitalium] >pirlSlB64249 tRNA (guanine-N1-)-methyltransferase (EC 2.1.1.31) - Mycoplasma genitalium (SGC3)	60	42
70	6	4740	5597	gplM33977l	Dpse\ry gene product [Drosophila pseudoobscura] >pirlSlA31946 xanthine dehydrogenase (EC 1.1.1.204) - fruit fly (Drosophila pseudoobscura)	60	42
101	8	3562	3326	gplM74329l	pcx gene product [Drosophila melanogaster] >gplM25662lDROPEC_1 pcx gene product [Drosophila melanogaster] {SUB 546-2483}	60	46
101	16	8112	7486	gplD64006l	hypothetical protein [Synechocystis sp.]	60	32
108	1	2	583	gplZ23080l	major vegetative sigma factor [Clostridium acetobutylicum] >pirlSlS34307 DNA-directed RNA polymerase (EC 2.7.7.6) sigma factor sigA - Clostridium acetobutylicum	60	38
124	1	917	3	gplX75568l	ICFG [unidentified] >pirlSlS48034 icfG protein - Synechocystis sp. (PCC 6803) >pirlSlS38573 ICFG protein - Synechocystis sp. (strain PCC6803)	60	33
130	1	111	719	gplL09228l	Bacillus subtilis spoVA to serA region. [Bacillus subtilis] >pirlSlS45549 hypothetical protein X7 - Bacillus subtilis	60	47
167	1	1	243	gplU35149l	putative glutamate and asparagine rich protein [Plasmodium chabaudi]	60	43
175	1	253	65	pirlSlA2966 6	keratin, 65K type II cytoskeletal - human >gplX05418lHSKER65A_1 keratin type II (AA1-215) [Homo sapiens] {SUB 1-215}	60	46
177	1	1077	199	gplU43739l	FtsW [Borrelia burgdorferi] >gplX96432lBBMRAYFTS_2 ftsW gene product [Borrelia burgdorferi] {SUB 1-27}	60	39
185	1	524	21	gplU17010l	NADH dehydrogenase, subunit 5 [Mitochondrion Allomyces macrogynus] >gplU41288lAMU41288_8 NADH dehydrogenase, subunit 5 [Mitochondrion Allomyces macrogynus]	60	45
570	1	2	532	gplJ03901l	Maize pyruvate,orthophosphate dikinase mRNA, complete cds. [Zea mays]	60	31

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

582	1	2	376	gp|U00008|	yojK [Escherichia coli]	60	40
586	1	265	2	gp|L45497|	ATP-dependent protease binding subunit [Haemophilus influenzae] >gp|U32767|HIU32767_7 ATP-dependent protease binding subunit [Haemophilus influenzae] >gp|U00078|HIU00078_6 ATP-dependent protease binding subunit [Haemophilus influenzae] >gp|U32713|HIU32713	60	37
606	1	368	3	gp|U23163|	OrfUU [Escherichia coli]	60	37
10	20	11015	10356	gp|D50064|	PgsA [Bacillus subtilis]	59	38
12	8	4396	4947	gp|U29580|	Escherichia coli K-12 genome; approximately 62 minute region. [Escherichia coli]	59	38
17	4	2628	2068	gp|U00021|	u0247g [Mycobacterium leprae]	59	42
19	13	10882	10502	gp|M88489|	nonamer binding protein [Mus musculus] >pir|S|JC4236 V(D)J recombinational signal sequence-dependent DNA joining protein 2 - mouse	59	38
19	19	14336	14151	gp|L49336|	tryptophanyl-tRNA synthetase [Clostridium longisporum]	59	40
19	20	15203	14334	gp|L49336|	tryptophanyl-tRNA synthetase [Clostridium longisporum]	59	39
30	2	1175	1402	gp|U18792|	glutamine-dependent carbamoyl phosphate synthase [Babesia bovis]	59	44
30	7	3181	4278	gp|Z23080|	primase [Clostridium acetobutylicum] >pir|S|S34306 DNA primase - Clostridium acetobutylicum	59	38
30	8	4242	4493	gp|U09825|	acid finger protein [Homo sapiens] >gp|U09825|HSU09825_1 acid finger protein [Homo sapiens]	59	37
34	4	2138	2740	gp|L44663|	hypothetical protein (GB:U00019_14) [Haemophilus influenzae] >gp|U32687|HIU32687_8 hypothetical protein (GB:U00019_14) [Haemophilus influenzae] >gp|U00069|HIU00069_17 hypothetical protein (GB:U00019_14) [Haemophilus influenzae] >gp|U32796|HIU32796_11 oxid	59	37
35	29	17973	17338	gp|Z34469|	class II metallothionein with homology to wheat Ec [Zea mays] >gp|U10696|ZMU10696_1 Ec metallothionein class II protein [Zea mays] >pir|S|S47158 metallothionein II - maize	59	54
42	7	3528	4172	gp|M33293|	E.coli primosomal protein n' (priA) gene, complete cds, and cytR gene, 5' end. [Escherichia coli]	59	43

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

48	12	9614	8892	gp|L44958|	Holliday junction DNA helicase [Haemophilus influenzae] >gp|U32716|HIU32716_14 Holliday junction DNA helicase [Haemophilus influenzae] >gp|U00072|HIU00072_43 Holliday junction DNA helicase [Haemophilus influenzae] >gp|U32825|HIU32825_13 Holliday junction	59	40
49	9	7040	6501	gp|D26185|	unknown [Bacillus subtilis]	59	37
53	3	1302	424	gp|L44984|	hypothetical protein (SP:P32049) [Haemophilus influenzae] >gp|U32719|HIU32719_1 hypothetical protein (SP:P32049) [Haemophilus influenzae] >gp|U00072|HIU00072_69 hypothetical protein (SP:P32049) [Haemophilus influenzae] >gp|U32828|HIU32828_1 SAM-dependent	59	41
63	25	14385	13177	gp|Y00544|	Eschericia coli gyrA gene, orfX and orfY'. [Escherichia coli] >gp|M87509|ECOUBIG_1 ubiquinone synthesis-related protein [Escherichia coli] >pir|S|A47682 2-octaprenyl-3-methyl-5-hydroxy-6-methoxy-1,4-benzoquinone methyltransferase, UbiG - Escherichia coli	59	40
63	32	19064	18459	gp|L19521|	Synechococcus sp. four ORFs, three complete, and one 3' end. [Synechococcus sp.]	59	41
64	5	3262	2102	gp|D50617|	Cytoplasmic phenylalanyl-tRNA synthetase beta chain [Mitochondrion Saccharomyces cerevisiae] >gp|D50617|YSCCHRVIN_47 Cytoplasmic phenylalanyl-tRNA synthetase beta chain [Saccharomyces cerevisiae] >pir|S|YFBYAC phenylalanine--tRNA ligase (EC 6.1.1.20) beta	59	40
70	26	16844	17113	gp|U24569|	sigma factor [Pseudomonas aeruginosa] >gp|L36379|PSEALGT_2 alternative sigma factor [Pseudomonas aeruginosa] >gp|L14760|PSEALGUA_1 algU gene product [Pseudomonas aeruginosa] >gp|L14761|PSEALGUB_1 algU gene product [Pseudomonas aeruginosa] >gp|U49151|PAU49	59	40
76	3	3089	2139	pir|S|B49696	smf - Escherichia coli	59	37
79	8	3348	3995	gp|L40501|	flagellar MS-ring protein [Borrelia burgdorferi]	59	33
79	10	5190	6212	gp|L40502|	flagellar export protein [Borrelia burgdorferi] >gp|L76303|BORFTSA_12 flagellar export apparatus [Borrelia	59	29

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					burgdorferi]		
85	8	3841	3398	gp|M93239|	transmembrane protein [Escherichia coli] >pir|S|C45313 putrescine transport protein potH - Escherichia coli	59	36
92	1	1	612	gp|L25603|	protease [Treponema denticola]	59	33
94	1	825	118	gp|X55034|	UDP-MurNac-pentapeptide presynthetase (AA -20 to 432) [Escherichia coli] >gp|X15432|ECMURF_1 UDP-MurNAC-pentapeptide presynthetase (AA -20 to 432) [Escherichia coli] >gp|D10483|ECO110K_66 UDP-N-acetylmuramoylalanyl-D-glutamyl-2, 6-diaminopimelate-D-alanyl	59	43
106	2	656	1174	gp|J02714|	Chicken embryonic myosin heavy chain gene, complete cds. [Gallus gallus] >pir|S|B24124 myosin heavy chain, EFW1 - chicken (fragment) {SUB 1-168}	59	20
114	1	3	272	gp|D64000|	hypothetical protein [Synechocystis sp.] >gp|U38915|SSU38915_1 LytB [Synechocystis sp.] {SUB 28-406}	59	35
114	4	995	1462	gp|U15180|	htrA [Mycobacterium leprae]	59	45
122	3	1795	2625	gp|X56678|	dciAE gene product [Bacillus subtilis]	59	41
136	5	2556	3410	gp|L46086|	H. influenzae predicted coding region HI1454 [Haemophilus influenzae] >gp|U32823|HIU32823_13 H. influenzae predicted coding region HI1454 [Haemophilus influenzae] >gp|U00084|HIU00084_49 H. influenzae predicted coding region HI1454 [Haemophilus influenzae]	59	37
138	1	609	382	gp|D26185|	unknown [Bacillus subtilis]	59	35
142	1	231	1328	gp|U32164|	NAD(P)H-dependent dihydroxyacetone-phosphate reductase [Bacillus subtilis]	59	41
202	1	722	183	gp|L44828|	hypothetical protein (SP:P30143) [Haemophilus influenzae] >gp|U32703|HIU32703_7 hypothetical protein (SP:P30143) [Haemophilus influenzae] >gp|U00070|HIU00070_87 hypothetical protein (SP:P30143) [Haemophilus influenzae] >gp|U32812|HIU32812_7 amino acid per	59	40
595	1	1	378	gp|M85240|	flagellar protein [Escherichia coli]	59	30
4	5	3965	4324	gp|X65964|	nestin [Homo sapiens] >pir|S|S21424 nestin - human	58	36

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

5	9	6219	6545	gp|Z35117|	alpha 1,4-glucan phosphorylase type H [Vicia faba]	58	52
9	5	4201	5055	gp|X54933|	SPase I [Salmonella typhimurium] >pir|S|S12020 signal peptidase I - Salmonella typhimurium	58	41
10	1	2596	1991	gp|X82174|	deoxyribose-phosphate aldolase [Bacillus subtilis] >pir|S|S49455 deoxyribose-phosphate aldolase (EC 4.1.2.4) - Bacillus subtilis	58	44
18	4	5354	3198	gp|U43739|	Borrelia burgdorferi fesmid clone 31, complete sequence. [Borrelia burgdorferi]	58	37
25	2	1075	149	gp|L14437|	flagellar hook-filament junction protein [Bacillus subtilis]	58	34
26	5	3419	3730	gp|D26185|	unknown [Bacillus subtilis]	58	36
28	19	8985	8803	gp|V00141|	Cauliflower mosaic virus genome. [Cauliflower mosaic virus] >pir|S|QQCV6S hypothetical protein 6 - cauliflower mosaic virus (strain Strasbourg)	58	58
29	20	9207	8524	gp|U29399|	major outer sheath protein [Treponema denticola]	58	45
41	2	547	945	gp|M21994|	E.coli cysK gene, 3' end, ptsH, ptsI, and crr phototransferase system genes, complete cds. [Escherichia coli] >gp|J02796|ECOPTSHI_2 ptsI gene product [Escherichia coli] >pir|S|WQECPI phosphotransferase system enzyme I (EC 2.7.3.9) - Escherichia coli	58	37
45	10	8472	9128	pir|S|B55053	endothelial monocyte-activating protein II precursor - human	58	27
48	21	13329	12793	gp|U18997|	Escherichia coli K-12 chromosomal region from 67.4 to 76.0 minutes. [Escherichia coli] >gp|U15661|ECU15661_1 HhoA [Escherichia coli] >gp|U32495|ECU32495_1 DegQ [Escherichia coli]	58	35
65	4	451	1836	gp|L29189|	transmembrane receptor [Bacillus subtilis] >pir|S|C54078 chemotaxis transducer homolog TLPA - Bacillus subtilis	58	30
79	20	11050	10535	gp|X54459|	traI gene product [Escherichia coli] >pir|S|S23001 traI protein - Escherichia coli plasmid RP4 >pir|S|C36042 traI protein - Escherichia coli plasmid RP4 (fragment) {SUB 1-21}	58	35
85	6	2589	2275	gp|M64519|	transport protein [Escherichia coli] >pir|S|D40840 spermidine/putrescine transport protein D - Escherichia coli	58	38
88	5	2328	3020	gp|L47617|	patH gene product [Vibrio harveyi]	58	37

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

100	17	4026	4235	gp|S53251|	orf V1...orf C4 [tomato leaf curl virus TLCV, Australian isolate, Genomic Complete, 6 genes, 2766 nt]. [Unknown.] >pir|S|JQ1888 AL2 protein - tomato yellow leaf curl virus (strain Australia)	58	47
100	22	7823	7101	gp|M77039|	E.coli MsbB protein gene, complete cds. [Escherichia coli] >pir|S|B42608 OrfU upstream of msbB - Escherichia coli (fragment)	58	42
101	33	19952	20701	gp|M57692|	membrane protein for maltose transport [Thermoanaerobacterium thermosulfurigenes] >gp|M57692|TTPULSA_4 membrane protein for maltose transport [Thermoanaerobacterium thermosulfurigenes] >pir|S|S37704 amyD protein - Thermoanaerobacterium thermosulfurigenes	58	33
104	11	4746	5162	gp|U28377|	Escherichia coli K-12 genome; approximately 65 to 68 minutes. [Escherichia coli]	58	49
112	12	7865	8896	gp|M26929|	D-2-hydroxyisocaproate dehydrogenase [Lactobacillus casei] >gp|A14931|A14931_1 D-2-Hydroxyisopransaeure-dehydrogenase [Lactobacillus casei] >pir|S|DELBC D-2-hydroxyisocaproate dehydrogenase (EC 1.1.1.-) - Lactobacillus casei	58	38
114	3	369	1082	gp|U18997|	Escherichia coli K-12 chromosomal region from 67.4 to 76.0 minutes. [Escherichia coli] >gp|U15661|ECU15661_2 HhoB [Escherichia coli] >gp|U32495|ECU32495_2 DegS [Escherichia coli]	58	41
122	6	2873	3442	gp|X56678|	dciAE gene product [Bacillus subtilis]	58	30
128	4	4092	3661	gp|L44753|	hypothetical protein (GB:U14003_130) [Haemophilus influenzae] >gp|U32696|HIU32696_3 hypothetical protein (GB:U14003_130) [Haemophilus influenzae] >gp|U00070|HIU00070_14 hypothetical protein (GB:U14003_130) [Haemophilus influenzae] >gp|U32805|HIU32805_3 in	58	34
134	3	1234	1025	gp|U33210|	methyl accepting chemotaxis homolog [Treponema denticola]	58	39
136	3	1970	1038	gp|M93419|	aspartokinase II [Bacillus sp.] >pir|S|A48946 aspartate kinase (EC 2.7.2.4) II precursor - Bacillus sp. (strain MGA3)	58	41
141	8	4280	4942	gp|U23163|	OrfUU [Escherichia coli]	58	43

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

142	2	1395	3965	gp|X59543|	M1 subunit of ribonucleotide reductase [Homo sapiens] >gp|X59617|HSRR1LS_1 large subunit ribonucleotide reductase [Homo sapiens] >pir|S|S16680 ribonucleoside-diphosphate reductase (EC 1.17.4.1) chain M1 - human	58	41
167	2	732	1049	gp|X83598|	elongation factor Ts [Thermus aquaticus thermophilus] >pir|S|S51095 elongation factor Ts - Thermus aquaticus	58	30
179	1	3	452	gp|U24253|	folylpolyglutamate synthetase [Homo sapiens]	58	37
566	4	995	1270	gp|X15041|	type VI collagen alpha-2 subunit preprotein [Gallus gallus] >gp|X56659|GDCOL6A2G_1 type VI collagen subunit alpha2 [Gallus gallus] >pir|S|S04111 collagen alpha 2(VI) chain long form precursor - chicken >gp|X56595|GGCOLVIA_1 type VI collagen alpha-2 subuni	58	41
634	1	12	467	gp|L44753|	hypothetical protein (GB:U14003_130) [Haemophilus influenzae] >gp|U32696|HIU32696_3 hypothetical protein (GB:U14003_130) [Haemophilus influenzae] >gp|U00070|HIU00070_14 hypothetical protein (GB:U14003_130) [Haemophilus influenzae] >gp|U32805|HIU32805_3 in	58	32
1	11	3440	4156	gp|D83536|	proline-tRNA ligase [Escherichia coli]	57	38
1	17	8526	8290	gp|M32474|	Rattus norvegicus carcinoembryonic antigen-related protein (CGM1) mRNA, complete cds. [Rattus norvegicus] >pir|S|A35364 carcinoembryonic antigen-related protein (clone rnCGM1) - rat	57	37
8	46	27159	27719	gp|U25996|	DnaJ [Haemophilus ducreyi]	57	44
15	2	359	862	gp|M35106|	Rat heart-derived c-ros-1 proto-oncogene mRNA, complete cds. [Rattus norvegicus]	57	30
15	4	1688	2896	gp|X15981|	E. coli sbcC gene (ORF-45) for SbcC. [Escherichia coli] >pir|S|JS0349 hypothetical 45K protein (sbcC 5' region) - Escherichia coli	57	31
16	3	2001	1369	gp|Z15056|	murE gene product [Bacillus subtilis] >pir|S|B47691 UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase (EC 6.3.2.13) - Bacillus subtilis	57	34
16	47	27505	28170	pir|S|S19739	integral membrane protein - Rhodobacter capsulatus	57	31

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

23	5	1554	2603	gplM63007l	mutS gene product [Azotobacter vinelandii] >pirlSlA53296 DNA mismatch repair protein MutS - Azotobacter vinelandii	57	34
26	26	14567	15556	gplU39691l	methylgalactoside permease ATP-binding protein [Mycoplasma genitalium] >pirlSlB64213 methylgalactoside permease ATP-binding protein homolog - Mycoplasma genitalium (SGC3) >gplU02149lMGU02149_1 Mycoplasma genitalium random genomic clone sc8a, partial cds.	57	36
26	29	16954	17610	gplU39691l	hypothetical protein (SP:P32720) [Mycoplasma genitalium] >pirlSlD64213 hypothetical protein homolog MG121 - Mycoplasma genitalium (SGC3)	57	31
29	23	10614	10126	gplX64558l	aprF gene product [Pseudomonas aeruginosa] >pirlSlS26698 aprF protein - Pseudomonas aeruginosa	57	34
40	3	1163	1582	gplM77039l	E.coli MsbB protein gene, complete cds. [Escherichia coli] >pirlSlB42608 OrfU upstream of msbB - Escherichia coli (fragment)	57	40
42	23	15434	16132	gplX72695l	RNA polymerase, beta' subunit (prime) [Thermotoga maritima] >pirlSlS41467 DNA-directed RNA polymerase (EC 2.7.7.6) beta' chain (prime) - Thermotoga maritima	57	34
44	23	13698	13420	gplU29134l	transketolase [Xanthobacter flavus] >gplU29134lXFU29134_1 transketolase [Xanthobacter flavus]	57	42
55	12	12176	10743	gplU33007l	D9461.18p [Saccharomyces cerevisiae]	57	40
63	34	21861l	20098	pirlSlR3EC1	ribosomal protein S1 - Escherichia coli >pirlSlS29161 ribosomal protein S1 - Escherichia coli (fragment) {SUB 1-20} >gplX00785lECRPSA01_2 E. coli rpsA operon leader sequence. [Escherichia coli] {SUB 1-21}	57	38
64	3	884	1783	gplL46284l	hypothetical protein (GB:D26185_99) [Haemophilus influenzae] >gplU32838lHIU32838_6 hypothetical protein (GB:D26185_99) [Haemophilus influenzae] >gplU00086lHIU00086_18 hypothetical protein (GB:D26185_99) [Haemophilus influenzae] >gplU32785lHIU32785_1 methy	57	40
70	25	15895	16557	gplX56234l	cysteine-tRNA ligase [Escherichia coli]	57	42
74	16	9350	9976	gplU12289l	prolipoprotein diacylglyceryl transferase [Escherichia coli]	57	42

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					>pirlS|A56149 prolipoprotein diacylglyceryl transferase (EC 2.-.-.-) - Escherichia coli		
76	4	3657	3022	gp|U43739|	Borrelia burgdorferi fesmid clone 31, complete sequence. [Borrelia burgdorferi] >gp|L76303|BORFTSA_3 Borrelia burgdorferi ftsA gene, 3' end of cds, ftsZ, orf230, smf, hslVU, flgBCE, fliEFGHI, flbABC genes, complete cds. [Borrelia burgdorferi] >gp|X96685|B	57	28
81	12	9993	11123	gp|U49269|	unknown [Moraxella catarrhalis]	57	32
81	21	18989	17592	gp|M63176|	helicase [Staphylococcus aureus] >pirlS|S27667 DNA helicase pcrA - Staphylococcus aureus >pirlS|S39923 pcrA protein - Staphylococcus aureus	57	37
100	21	5706	7010	gp|X61190|	L.delbrueckii nifS-like gene (partial). [Lactobacillus delbrueckii] >pirlS|S16047 nifS protein homolog - Lactobacillus delbrueckii	57	41
101	21	11344	12489	gp|U02965|	unknown [Escherichia coli]	57	46
101	26	16171	16473	gp|L45564|	H. influenzae predicted coding region HI0926 [Haemophilus influenzae] >gp|U32774|HIU32774_6 H. influenzae predicted coding region HI0926 [Haemophilus influenzae] >gp|U00078|HIU00078_71 H. influenzae predicted coding region HI0926 [Haemophilus influenzae]	57	35
102	6	3868	3302	gp|L13867|	penicillin-binding protein 1A [Pseudomonas aeruginosa]	57	34
102	7	4929	4129	gp|U07173|	TagE [Vibrio cholerae] >gp|U39068|VCU39068_4 Vibrio cholerae pathogenicity island, partial and complete cds. [Vibrio cholerae] >pirlS|JC2569 tagE protein - Vibrio cholerae (strain 0395)	57	39
103	1	182	703	gp|D26185|	unknown [Bacillus subtilis]	57	42
104	20	8066	8869	gp|U36840|	recN gene product [Escherichia coli]	57	33
121	32	17961	18137	gp|M27082|	GCN2 gene product [Saccharomyces cerevisiae] >pirlS|OKBYN2 protein kinase GCN2 (EC 2.7.1.-) - yeast (Saccharomyces cerevisiae) >gp|U51030|YSCD9954_16 Protein kinase, phosphorylates the alpha subunit of eIF-2 (Swiss prot. accession number P15442) [Saccharo	57	47
121	33	19184	18273	gp|U00006|	E. coli chromosomal region from 89.2 to 92.8 minutes. [Escherichia coli]	57	29

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

128	5	4206	4835	gp|U09711|	flagella switch protein [Borrelia burgdorferi] >gp|L76303|BORFTSA_11 flagellar switch protein [Borrelia burgdorferi]	57	40
141	4	2425	2622	gp|L21195|	serotonin 5-HT7 receptor protein [Homo sapiens]	57	33
144	1	772	62	pir|S|A60332	80K antigen - Lyme disease spirochete >gp|X81514|BBP831001_1 p83/100 gene product [Borrelia burgdorferi] {SUB 381-475}	57	46
150	5	3378	4838	gp|U00019|	B2235_C2_195 [Mycobacterium leprae]	57	40
183	1	1	555	gp|X51738|	flgJ gene product (5 AA) [Salmonella typhimurium] >pir|S|SMEBH1 flagellar hook-associated protein 1 - Salmonella typhimurium >gp|M24466|STYFLGH_4 S.typhimurium flagellar L-ring (flgH), flagellar P-ring (flgI), and flagellar (flgJ) genes, complete cds. [Sa	57	28
16	7	5049	4462	gp|M24890|	esterase [Acinetobacter calcoaceticus]	56	41
16	38	22640	22410	gp|L14285|	signaling protein [Plasmid pCF10] >pir|S|C53309 prgY protein - Enterococcus faecalis plasmid pCF10	56	35
19	14	13098	10726	gp|M24278|	lig gene product [Escherichia coli]	56	38
21	6	4067	4681	gp|L77246|	yplQ gene product [Bacillus subtilis]	56	42
22	15	9726	10397	gp|X59399|	AdgA protein [Rhodobacter capsulatus] >pir|S|S15555 adgA protein - Rhodobacter capsulatus	56	39
29	10	4741	3581	gp|U00013|	nifS [Mycobacterium leprae]	56	39
33	1	1	366	gp|U39688|	hydroxymethylglutaryl-CoA reductase [Mycoplasma genitalium] >pir|S|D64209 hydroxymethylglutaryl-CoA reductase (NADPH) homolog - Mycoplasma genitalium (SGC3)	56	30
41	4	1447	1704	gp|U00006|	E. coli chromosomal region from 89.2 to 92.8 minutes. [Escherichia coli]	56	33
42	12	6057	6617	gp|M34995|	B.subtilis minor sigma-37 factor of RNA polymerase (rpoF, sigB), complete cds. [Bacillus subtilis] >pir|S|A36131 hypothetical protein V (sigB 5' region) - Bacillus subtilis >gp|L35574|BACRSBU_5 sigma-B positive regulator [Bacillus subtilis] {SUB 1-21}	56	32
44	18	10412	11143	gp|M55917|	Spirochaeta aurantia anthranilate synthase component I (trpE)	56	34

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					gene, complete cds, and ORF's 1,2 and 3. [Spirochaeta aurantia]		
47	16	10115	9225	gp|X81642|	orf gene product [Wolinella succinogenes] >pir|S|S50154 hypothetical protein - Wolinella succinogenes	56	34
48	1	765	4	gp|U45285|	specific 116-kDa vacuolar proton pump subunit [Homo sapiens]	56	37
48	8	5573	6028	gp|M17643|	B.subtilis spoIIA locus sporulation genes. [Bacillus subtilis] >pir|S|A55646 stage II sporulation protein AA protein - Bacillus subtilis	56	25
49	11	8342	7806	gp|D26185|	unknown [Bacillus subtilis]	56	36
51	5	5745	6236	gp|X16817|	Klebsiella pneumoniae gyrA gene for DNA gyrase subunit A (EC 5.99.1.3). [Klebsiella pneumoniae]	56	38
61	9	3590	4795	gp|D26185|	unknown [Bacillus subtilis]	56	40
63	24	13179	12289	gp|L10328|	f270 gene product [Escherichia coli]	56	34
64	28	11918	12559	gp|M64273|	transfer RNA-Met synthetase [Thermus thermophilus] >pir|S|SYTWMT methionine--tRNA ligase (EC 6.1.1.10) - Thermus aquaticus	56	36
74	11	6202	7569	gp|X75568|	ICFG [unidentified] >pir|S|S48034 icfG protein - Synechocystis sp. (PCC 6803) >pir|S|S38573 ICFG protein - Synechocystis sp. (strain PCC6803)	56	40
79	21	10889	11098	gp|M59449|	polypeptide chain-binding protein [Zea mays] >gp|M59449|MZEB70A_1 polypeptide chain-binding protein [Zea mays] >pir|S|JQ0966 immunoglobulin-binding protein homolog b70 - maize (fragment) {SUB 1-467}	56	34
85	5	2314	1505	gp|U25682|	Lpp38 [Pasteurella haemolytica]	56	36
87	5	3287	2613	gp|X73329|	uracil phosphoribosyltransferase [Lactococcus lactis]	56	34
101	13	6208	5834	gp|D26185|	replicative DNA helicase [Bacillus subtilis]	56	28
101	14	7059	6637	gp|U14003|	50S ribosomal subunit protein L9 [Escherichia coli] >gp|X04022|ECRPSFRI_4 E. coli genes rpsF, rpsR and rplI for ribosomal proteins S6, S18, L9. [Escherichia coli] >pir|S|S56428 50S ribosomal chain protein L9 - Escherichia coli	56	30
104	21	8701	9282	gp|M30297|	B.subtilis recombination and sporulation protein (recN, spoIVB) genes , complete cds, arginine hydroximate resistance (ahrC) gene, 3' end. [Bacillus subtilis] >pir|S|B35128 recN homolog - Bacillus	56	35

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					subtilis		
106	1	3	740	gp|L06148|	golgin-165 [Homo sapiens] >pir|S|JH0820 160K golgi antigen - human (fragment)	56	33
112	16	10511	11668	gp|U46542|	ScbA [Streptococcus crista]	56	36
115	2	1153	650	gp|X72382|	R.capsulatus nifR3 DNA. [Rhodobacter capsulatus] >pir|S|S34980 hypothetical protein (nifR3 5' region) - Rhodobacter capsulatus	56	43
150	2	3212	2412	gp|D28118|	DB1 [Homo sapiens] >gp|D28118|HUMDB1_1 DB1 [Homo sapiens] >pir|S|A53772 transcription factor DB1 - human	56	39
150	11	10373	9024	gp|L15202|	Bacillus subtilis comE operon encoding ORF1, ORF2, ORF3 and Reverse-ORF genes, complete cds. [Bacillus subtilis] >pir|S|S39864 ComE ORF2 - Bacillus subtilis	56	39
161	4	1128	730	gp|S51224|	phytochrome [Ceratodon purpureus] >pir|S|S27396 phytochrome - moss (Ceratodon purpureus) >pir|S|S12966 phytochrome - moss (Ceratodon purpureus) {SUB 49-539}	56	37
697	1	213	7	gp|M60177|	enterobactin [Escherichia coli] >gp|M60177|ECOENTF_1 enterobactin [Escherichia coli] >pir|S|YGECEF enterochelin synthetase (EC 6.-.-.-) component F - Escherichia coli >gp|M17354|ECOENTFA_1 E.coli entF gene encoding serine activating enzyme, 3' end. [Esche	56	36
733	1	455	243	gp|X14309|	Murine mRNA for 4F2 antigen heavy chain. [Mus musculus] >pir|S|S03600 cell surface antigen 4F2 heavy chain - mouse	56	40
1	18	9516	8524	gp|X96983|	hypothetical protein [Bacillus subtilis]	55	40
4	2	2681	1215	gp|Z68195|	Gcd6p [Saccharomyces cerevisiae] >gp|L07115|YSCGCDA_2 guanine nucleotide exchange factor, eIF-2B, delta subunit [Saccharomyces cerevisiae] >gp|Z68194|SC8142A_12 Gcd6p [Saccharomyces cerevisiae] >pir|S|A48156 translation regulator GCD6 - yeast (Saccharomyc	55	27
4	9	6375	7040	gp|X87113|	DnaJ protein [Agrobacterium tumefaciens]	55	47
5	4	2097	2438	gp|L45445|	H. influenzae predicted coding region HI0807 [Haemophilus influenzae] >gp|U32763|HIU32763_2 H. influenzae predicted coding region HI0807 [Haemophilus influenzae] >gp|U00077|HIU00077_66 H. influenzae predicted coding region	55	36

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					HI0807 [Haemophilus influenzae]		
8	15	7544	8125	gpl|L45104|	oxygen-independent coproporphyrinogen III oxidase [Haemophilus influenzae] >gpl|U32729|HIU32729_6 oxygen-independent coproporphyrinogen III oxidase [Haemophilus influenzae] >gpl|U00073|HIU00073_85 oxygen-independent coproporphyrinogen III oxidase [Haemophil	55	32
16	10	7078	6437	gpl|U39698|	M. genitalium predicted coding region MG207 [Mycoplasma genitalium] >pir||S|H64222 hypothetical protein MG207 - Mycoplasma genitalium (SGC3)	55	28
16	54	33148	32180	gpl|X84019|	orf3 gene product [Zymomonas mobilis] >gpl|X84019|ZMDNAGRP_3 orf3 gene product [Zymomonas mobilis]	55	24
19	4	2825	3775	gpl|F14533|	histidyl tRNA synthetase [Sus scrofa]	55	39
21	15	9866	10165	gpl|L17342|	Valyl-tRNA synthetase [Haemophilus parainfluenzae]	55	34
22	1	105	1673	gpl|Z46812|	ZK675.1 [Caenorhabditis elegans] >gpl|Z46812|CEZK675_1 ZK675.1 [Caenorhabditis elegans]	55	31
27	6	2720	3091	gpl|J04243|	unknown protein [Salmonella typhimurium] >pir||S|C32890 hypothetical protein (prfA 3' region) - Salmonella typhimurium (fragment)	55	27
28	1	808	137	gpl|X95575|	cytochrome oxidase subunit 2 [Mitochondrion Chorthippus parallelus]	55	44
44	19	11435	12082	gpl|L45882|	H. influenzae predicted coding region HI1248 [Haemophilus influenzae] >gpl|U32805|HIU32805_1 H. influenzae predicted coding region HI1248 [Haemophilus influenzae] >gpl|U00082|HIU00082_17 H. influenzae predicted coding region HI1248 [Haemophilus influenzae]	55	38
44	21	12478	13056	gpl|U39722|	M. genitalium predicted coding region MG372 [Mycoplasma genitalium] >pir||S|B64241 hypothetical protein MG372 - Mycoplasma genitalium (SGC3)	55	31
51	1	2009	3241	gpl|D26185|	recombination protein [Bacillus subtilis] >gpl|X02369|BSORIC_5 Bacillus subtilis oriC region. [Bacillus subtilis] {SUB 48-370}	55	31
53	7	3564	3352	gpl|L47709|	ypiA gene product [Bacillus subtilis]	55	27

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

54	25	7958	8197	gp|V01146|	Genome of bacteriophage T7. [Bacteriophage T7] >pir|S|JVBPB7 DNA maturase B - phage T7 >pir|S|S42338 gene 19 protein - phage T7	55	41
66	1	979	2	gp|Z48001|	orf3 gene product [Thermus aquaticus thermophilus] >pir|S|S52278 hypothetical protein 3 - Thermus aquaticus	55	41
70	22	15248	15571	pir|S|F53402	cysteine--tRNA ligase (EC 6.1.1.16) - Bacillus stearothermophilus (fragment)	55	31
80	4	1771	1226	gp|U09005|	HflC [Vibrio parahaemolyticus]	55	39
86	1	232	468	gp|Z21970|	54CP [Chloroplast Arabidopsis thaliana] >pir|S|S36637 signal recognition particle 54CP protein precursor - Arabidopsis thaliana	55	40
101	31	19067	18792	gp|L23195|	cytoplasmic dynein heavy chian [Drosophila melanogaster] >pir|S|A54794 dynein heavy chain, cytoplasmic - fruit fly (Drosophila melanogaster) >gp|L25122|DRODYNEINH_1 dynein heavy chain [Drosophila melanogaster] {SUB 1877-1998}	55	44
113	1	1	1668	gp|X68309|	ERCC3 gene product [Drosophila melanogaster] >pir|S|S26719 ERCC3 protein - fruit fly (Drosophila melanogaster)	55	37
122	5	2451	2308	gp|M64780|	agrin [Rattus norvegicus]	55	44
135	6	3068	3406	gp|L45521|	D-alanine permease [Haemophilus influenzae] >gp|U32770|HIU32770_3 D-alanine permease [Haemophilus influenzae] >gp|U00078|HIU00078_28 D-alanine permease [Haemophilus influenzae] >gp|U32716|HIU32716_5 alanine permease [Haemophilus influenzae] >pir|S|H64099	55	35
175	4	2312	1494	gp|D26134|	297 amino acids peptide, unknown function [Pseudomonas aeruginosa]	55	44
183	2	699	917	gp|J01917|	52,55kD protein [Human adenovirus type 2] >gp|J01917|ADRCG_13 52,55kD protein [Human adenovirus type 2] >pir|S|WMAD52 late L1 52K protein - human adenovirus 2 >gp|M73260|ADRCOMPGEN_1 Mastadenovirus h5 gene, complete genome. [Mastadenovirus h5] {SUB 173-41	55	38
563	1	1	501	gp|M24537|	GTP-binding protein [Bacillus subtilis] >pir|S|B32804 GTP-binding protein, spo0B 3'-region - Bacillus subtilis >gp|K02666|BACSPO0B2_2 Bacillus subtilis spo0B early	55	37

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					sporulation gene, complete cds. [Bacillus subtilis] (SUB 1-65)		
647	1	582	358	gp|U29399|	major outer sheath protein [Treponema denticola]	55	44
1	16	8237	7461	gp|M77039|	E.coli MsbB protein gene, complete cds. [Escherichia coli] >pir|S|B42608 OrfU upstream of msbB - Escherichia coli (fragment)	54	40
15	6	3330	3791	gp|X14298|	Human mRNA for dystrophin. [Homo sapiens] >gp|M86903|HUMDYST20_1 dystrophin gene product [Homo sapiens] (SUB 2980-3685) >gp|L05649|HUMDYSTR15_1 dystrophin [Homo sapiens] (SUB 2850-2979) >pir|S|S02109 dystrophin - human (fragment) (SUB 2305-2366) >gp|M2326	54	26
16	2	1371	1093	gp|L45769|	UDP-MurNac-tripeptide synthetase [Haemophilus influenzae] >gp|U32793|HIU32793_5 UDP-MurNac-tripeptide synthetase [Haemophilus influenzae] >gp|U00081|HIU00081_93 UDP-MurNac-tripeptide synthetase [Haemophilus influenzae] >gp|U32739|HIU32739_6 UDP-MurNac-tri	54	33
19	5	5650	3800	gp|U33210|	methyl accepting chemotaxis homolog [Treponema denticola]	54	30
19	22	17033	16728	gp|M35200|	M.xanthus frzG and frzF genes, complete cds. [Myxococcus xanthus] >pir|S|XYYZFG frzG protein - Myxococcus xanthus	54	29
26	12	7612	7229	gp|M60503|	profilaggrin [Homo sapiens]	54	40
36	21	16662	17306	gp|U28377|	Escherichia coli K-12 genome; approximately 65 to 68 minutes. [Escherichia coli]	54	20
37	8	2912	3670	gp|U39695|	preprotein translocase secY subunit [Mycoplasma genitalium] >pir|S|H64218 preprotein translocase secY - Mycoplasma genitalium (SGC3)	54	31
44	22	13047	13349	gp|U39722|	M. genitalium predicted coding region MG372 [Mycoplasma genitalium] >pir|S|B64241 hypothetical protein MG372 - Mycoplasma genitalium (SGC3)	54	35
47	29	15854	16273	gp|D12681|	esterase [Bacillus stearothermophilus] >pir|S|JC1374 carboxylesterase (EC 3.1.1.1) - Bacillus stearothermophilus (strain IFO 12550)	54	38
50	1	1	231	gp|J04740|	S.acidocaldanis ATP synthase (non-FoF1) membrane subunit P (atpP) gene, complete cds. [Sulfolobus acidocaldarius]	54	33

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					>pir|S|A33351 H+-transporting ATP synthase (EC 3.6.1.34) proteolipid chain - Sulfolobus acidocaldarius		
63	20	8698	7433	gp|U00039|	xylulokinase [Escherichia coli] >gp|K01996|ECOXYLABA_2 xylB gene product [Escherichia coli] >gp|X04691|ECXYLIK_2 E. coli genes for xylose isomerase and xylulose kinase. [Escherichia coli] >pir|S|KIECXY xylulokinase (EC 2.7.1.17) - Escherichia coli	54	26
71	2	1292	567	gp|U35369|	D,D-carboxypeptidase [Enterococcus faecalis]	54	38
100	37	15457	17007	gp|X91047|	hook-associated protein 2 [Xenorhabdus nematophilus]	54	22
101	34	20826	21446	gp|X66092|	C.perfringens ORF for putative membrane transport protein. [Clostridium perfringens] >pir|S|A56641 probable membrane transport protein - Clostridium perfringens	54	31
112	2	1870	914	gp|D64006|	hypothetical protein [Synechocystis sp.]	54	33
117	6	4154	4732	gp|M68971|	hexokinase type II [Rattus norvegicus] >pir|S|S15885 hexokinase (EC 2.7.1.1) II precursor - rat >gp|M68972|RATHKIINAH_1 hexokinase type II [Rattus norvegicus] {SUB 402-917} >gp|Z46367|HSHKEX16_1 hexokinase II [Homo sapiens] {SUB 741-791} >pir|S|S52114 typ	54	35
121	7	2265	3446	gp|Z25865|	SpoVD [Bacillus subtilis] >pir|S|S43863 SpoVD protein - Bacillus subtilis {SUB 1-589} >gp|L09703|BACPBPSPOV_4 penicillin-binding protein [Bacillus subtilis] {SUB 1-69} >gp|Z15056|BSSPOG_1 SpoVD [Bacillus subtilis] {SUB 595-645}	54	34
126	2	257	1354	gp|X96685|	phospho-N-acetylmuramoyl-pentapeptide- transferase [Borrelia burgdorferi] >gp|X96432|BBMRAYFTS_1 phospho-N-acetylmuramoyl-pentapeptide- transferase [Borrelia burgdorferi]	54	29
136	4	1964	2590	gp|L09228|	Bacillus subtilis spoVA to serA region. [Bacillus subtilis] >pir|S|S45556 hypothetical protein X14 - Bacillus subtilis	54	35
154	1	887	3	gp|X75439|	isoleucyl tRNA synthetase [Staphylococcus aureus]	54	33
566	1	329	751	gp|X69292|	smooth muscle mysosin heavy chain [Homo sapiens]	54	35
677	1	3	506	gp|U29399|	major outer sheath protein [Treponema denticola]	54	35
1	12	4378	5430	gp|L12684|	Pseudomonas putida recA protein gene, complete cds.	53	42

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					[Pseudomonas putida]		
12	4	3871	2531	gp|X73124|	ipa-67d gene product [Bacillus subtilis] >pir|S|S39722 hypothetical protein - Bacillus subtilis	53	38
13	1	666	4	gp|X56958|	ankyrin (brank-2) [Homo sapiens]	53	36
13	11	6925	6701	gp|U17246|	Belongs to the ATP-binding transport protein family (ABC transporters) [Saccharomyces cerevisiae] >pir|S|S51433 MDL1 protein - yeast (Saccharomyces cerevisiae)	53	30
16	19	12790	12350	gp|L22167|	LZIP-1 and LZIP-2 [Mus musculus]	53	29
26	6	4160	4489	gp|X79075|	orf 208 gene product [Coxiella burnetii] >pir|S|S44297 hypothetical protein 208 - Coxiella burnetii	53	38
32	4	3007	3594	gp|U35673|	Borrelia burgdorferi OrfR gene, partial cds, and S20, HBbu, OrfH and Rho genes, complete cds. [Borrelia burgdorferi]	53	28
35	28	17196	16795	gp|M81168|	rubredoxin oxidoreductase [Desulfovibrio vulgaris] >gp|M28848|DVURUBRBO_1 rubredoxin oxidoreductase [Desulfovibrio vulgaris] >pir|S|RDDVBX rubredoxin--NAD+ reductase (EC 1.18.1.1) - Desulfovibrio vulgaris	53	41
36	19	13115	13906	gp|M29701|	S.typhimurium polymerase III polymerase subunit gene, complete cds. [Salmonella typhimurium] >pir|S|A45915 DNA-directed DNA polymerase (EC 2.7.7.7) III alpha chain - Salmonella typhimurium >gp|D49445|ECODNAE_1 DnaE, DNA polymerase III holoenzyme catalytic	53	39
42	6	3076	3669	gp|L44983|	primosomal protein replication factor [Haemophilus influenzae] >gp|U32718|HIU32718_10 primosomal protein replication factor [Haemophilus influenzae] >gp|U00072|HIU00072_68 primosomal protein replication factor [Haemophilus influenzae] >gp|U32827|HIU32827_	53	28
42	10	4662	5078	gp|X02499|	Rhodospirillum rubrum atp operon. [Rhodospirillum rubrum] >pir|S|S08579 hypothetical protein 2 - Rhodospirillum rubrum >gp|X02499|RRATP_4 Rhodospirillum rubrum atp operon. [Rhodospirillum rubrum] {SUB 592-811}	53	40
43	3	1637	2905	gp|X72491|	Ig kappa light chain (VJ) [Homo sapiens] >pir|S|S40381 Ig kappa chain - human	53	33

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

44	25	14821	14375	gp|U00013|	tkt [Mycobacterium leprae]	53	40
54	29	10385	8799	gp|U36427|	glutamate synthase small subunit gltD [Thiobacillus ferrooxidans]	53	38
55	14	13304	12729	gp|U15609|	flagellar switch protein [Treponema denticola] >gp|L36851|TRPFLIG_1 fliG gene product [Treponema denticola]	53	27
59	1	1597	359	gp|M77039|	E.coli MsbB protein gene, complete cds. [Escherichia coli] >pir|S|B42608 OrfU upstream of msbB - Escherichia coli (fragment)	53	38
70	14	10256	9798	gp|L26916|	Pseudomonas aeruginosa (rpoN) gene, complete cds; ORF1, complete cds; ORF2, complete cds. [Pseudomonas aeruginosa] >pir|S|C53373 hypothetical protein 2 (rpoN 3' region) - Pseudomonas aeruginosa	53	28
70	18	13714	12437	gp|U21853|	unknown [Anabaena sp.]	53	40
74	8	1548	3560	gp|D10280|	myosin heavy chain [Oryctolagus sp.] >pir|S|A38650 myosin heavy chain, embryonic smooth muscle - rabbit (fragment)	53	34
77	1	719	3	gp|U39688|	hydroxymethylglutaryl-CoA reductase [Mycoplasma genitalium] >pir|S|D64209 hydroxymethylglutaryl-CoA reductase (NADPH) homolog - Mycoplasma genitalium (SGC3)	53	33
79	14	7556	8062	pir|S|SJCHA	spectrin alpha chain, brain - chicken >gp|X14518|GGSPECA5_1 Chicken gene for spectrin alpha-chain 5' end. [Gallus gallus] {SUB 1-27}	53	20
81	11	8080	9603	gp|U49269|	amidase [Moraxella catarrhalis]	53	41
81	13	12604	11219	gp|D26185|	unknown [Bacillus subtilis] >gp|U02604|BSU02604_3 ORFY [Bacillus subtilis] {SUB 1-260}	53	36
88	6	2957	3565	gp|U18539|	FliY [Escherichia coli]	53	25
89	11	9584	9054	gp|X52905|	Escherichia coli betT, betI, betB and betA genes. [Escherichia coli] >pir|S|S15179 betT protein - Escherichia coli	53	30
100	39	17891	18439	gp|L19346|	Escherichia coli N-acetylmuramoyl-L-alanine amidase (amiB) gene, complete cds, DNA repair protein (mutL) gene, partial cds, and two unidentified cds's. [Escherichia coli] >gp|U14003|ECOUW93_80 yjeE gene product [Escherichia coli] >pir|S|S56393 hypothetica	53	39
100	41	18333	19052	gp|L45029|	H. influenzae predicted coding region HI0388 [Haemophilus	53	29

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					influenzae] >gp|U32722|HIU32722_10 H. influenzae predicted coding region HI0388 [Haemophilus influenzae] >gp|U00073|HIU00073_9 H. influenzae predicted coding region HI0388 [Haemophilus influenzae]		
121	6	1690	2394	gp|Z25865|	SpoVD [Bacillus subtilis] >pir|S|S43863 SpoVD protein - Bacillus subtilis {SUB 1-589} >gp|L09703|BACPBPSPOV_4 penicillin-binding protein [Bacillus subtilis] {SUB 1-69} >gp|Z15056|BSSPOG_1 SpoVD [Bacillus subtilis] {SUB 595-645}	53	34
125	1	1385	2176	gp|D50624|	Adenosine Deaminase [Streptomyces virginiae] >gp|D50624|STMVBRA1_1 adenosine deaminase [Streptomyces virginiae] {SUB 1-339}	53	37
128	3	4162	2798	gp|X73141|	hemolysin [Serpulina hyodysenteriae]	53	32
143	1	716	3	gp|L12722|	transcription factor IIIB 131kDa subunit [Saccharomyces cerevisiae] >pir|S|A47453 transcription factor IIIC chain TFC4 - yeast (Saccharomyces cerevisiae)	53	34
143	4	1066	1332	gp|X55489|	chromogranin B [Bos taurus]	53	32
150	1	2293	2730	gp|U35673|	Borrelia burgdorferi OrfR gene, partial cds, and S20, HBbu, OrfH and Rho genes, complete cds. [Borrelia burgdorferi]	53	28
150	3	2709	3380	gp|X07693|	protein p67 [Schizosaccharomyces pombe] >pir|S|A30185 nuclear protein nuc2+ - fission yeast (Schizosaccharomyces pombe)	53	35
156	2	58	795	gp|D45911|	hypothetical protein [Bacillus subtilis]	53	23
358	1	366	617	gp|X82209|	MN1 gene product [Homo sapiens]	53	53
670	1	327	151	gp|U29399|	major outer sheath protein [Treponema denticola]	53	46
2	14	11472	11702	gp|U19615|	LET 858 [Caenorhabditis elegans]	52	34
6	4	3991	3425	gp|X77091|	trkA gene product [Escherichia coli] >gp|U18997|ECOUW67_214 TrkA protein of the constitutive K+ transport system Trk [Escherichia coli] >gp|X52114|ECTRKAG_3 TrkA protein of the constitutive K+ -transport system Trk [Escherichia coli] >pir|S|S36252 trkA pr	52	29
10	9	6165	5398	gp|L47539|	cyclic AMP receptor protein [Haemophilus somnus]	52	30
11	14	6750	6917	gp|M75136|	Ictalurid herpesvirus 1 (channel catfish virus [CCV]), strain	52	39

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					auburn 1, complete genome. [Ictalurid herpesvirus 1] >gp|M75136|IH1CG_64 Ictalurid herpesvirus 1 (channel catfish virus (CCV)), strain auburn 1, complete genome. [Ictalurid herpesvirus 1] >pir|		
11	29	16664	17380	gp|L45923|	hypothetical protein (SP:P09171) [Haemophilus influenzae] >gp|U32809|HIU32809_2 hypothetical protein (SP:P09171) [Haemophilus influenzae] >gp|U00082|HIU00082_56 hypothetical protein (SP:P09171) [Haemophilus influenzae] >gp|U32754|HIU32754_5 pseudoU synthe	52	36
16	52	32135	30837	gp|L46186|	H. influenzae predicted coding region HI1555 [Haemophilus influenzae] >gp|U32830|HIU32830_11 H. influenzae predicted coding region HI1555 [Haemophilus influenzae] >gp|U00085|HIU00085_26 H. influenzae predicted coding region HI1555 [Haemophilus influenzae]	52	21
19	7	7656	5668	gp|U33210|	methyl accepting chemotaxis homolog [Treponema denticola]	52	31
21	2	1556	1332	gp|X74468|	late protein [Human papillomavirus type 15] >pir|S|S36478 late protein - human papillomavirus type 15 >gp|M96284|PPHL1DF_1 L1 gene product [Human papillomavirus type 15] (SUB 312-355)	52	41
23	1	205	2	gp|X70990|	sucrose synthase [Arabidopsis thaliana]	52	52
30	11	6103	6936	gp|U00802|	drebrin E2 [Homo sapiens] >gp|D17530|HUMDRE_1 drebrin E [Homo sapiens] >pir|S|JN0809 drebrin E (clone gDbh13) - human	52	25
42	3	1276	950	gp|L46248|	membrane associated ATPase [Haemophilus influenzae] >gp|U32835|HIU32835_5 membrane associated ATPase [Haemophilus influenzae] >gp|U00085|HIU00085_85 membrane associated ATPase [Haemophilus influenzae] >pir|S|A64133 membrane associated ATPase (cbiO) homolo	52	40
54	20	7489	6683	gp|L07958|	NAD-dependent methylenetetrahydrofolate dehydrogenase-methenyltetrahydrofolate cyclohydrolase [Drosophila melanogaster] >gp|S59910|S59910_1 NAD-dependent methylenetetrahydrofolate dehydrogenase [Drosophila melanogaster] >pir|S|S32562 methylenetetrahydrofo	52	33
55	9	8295	8795	gp|D37799|	orf4 [Bacillus subtilis] >gp|D37799|BACAMOKOOO_6 orf4 [Bacillus subtilis]	52	30

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

55	13	13002	12193	gp|U09711|	flagella switch protein [Borrelia burgdorferi] >gp|L76303|BORFTSA_11 flagellar switch protein [Borrelia burgdorferi]	52	29
63	31	18372	17776	gp|X90872|	gp25l2 gene product [Homo sapiens]	52	35
68	7	2743	3408	pir|S|JN0695	tributyltin chloride resistant protein - Alteromonas sp. (strain M-1)	52	35
70	20	14776	13976	gp|M31827|	Bacillus subtilis (clone lambda-BS1) cell division and sporulation protein (dds) gene, complete cds. [Bacillus subtilis] >pir|S|A43727 probable division initiation regulatory protein 1 - Bacillus subtilis >gp|M31800|BACDIV_1 B.subtilis division initiation	52	42
77	11	5580	5026	gp|L38252|	Acinetobacter lwoffii orf1 and esterase (est) genes, complete cds. [Acinetobacter lwoffii]	52	34
83	11	3711	4682	gp|Z54328|	unknown [Schizosaccharomyces pombe]	52	34
104	5	1730	1915	gp|S75665|	PSI [Drosophila] >gp|S75666|S75666_1 PSI {alternatively spliced} [Drosophila, Pre-mRNA, 3961 nt]. [Drosophila]	52	41
104	13	5305	5478	gp|U00024|	u0002q [Mycobacterium tuberculosis]	52	39
112	14	9864	9097	gp|M26130|	S. parasanguis adhesin (fimA), ORF1, ORF3, and ORF5 genes, complete cds. [Streptococcus parasanguis]	52	27
115	4	1919	1473	gp|Z56280|	carD gene product [Myxococcus xanthus]	52	35
170	4	3357	2311	gp|U43739|	FtsA [Borrelia burgdorferi] >gp|X96433|BBFTSWQA_3 ftsA gene product [Borrelia burgdorferi] {SUB 1-36}	52	32
182	2	1189	215	pir|S|B36871	orf34 5' of tgs - Escherichia coli	52	38
619	1	248	24	gp|Z17372|	M.smegmatis asd, ask-alpha, and ask-beta genes. [Mycobacterium smegmatis] >pir|S|S31804 hypothetical protein y - Mycobacterium smegmatis	52	27
631	1	403	621	gp|D64003|	hypothetical protein [Synechocystis sp.] >gp|D64003|SYCSLLE_50 hypothetical protein [Synechocystis sp.]	52	30
15	7	4152	5183	gp|L13973|	Gallus gallus EDT-soluble/130 kDa protein mRNA, complete cds. [Gallus gallus] >pir|S|A47168 cardiac morphogenesis protein ES/130 - chicken (fragment)	51	26

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

17	1	552	73	gp|U36561|	fus-like protein [Homo sapiens]	51	42
19	2	2349	556	gp|U00022|	u0308b [Mycobacterium leprae]	51	37
30	12	7372	8253	gp|X86903|	peptidylprolyl isomerase [Triticum aestivum] >pir|S|S55383 peptidylprolyl isomerase (EC 5.2.1.8) - wheat	51	29
30	14	9302	10168	gp|M31792|	E.coli MreB protein gene, 3' end, MreC protein gene, complete cds, and MreD protein gene, complete cds. [Escherichia coli] >gp|U18997|ECOUW67_180 mreC gene product [Escherichia coli] >pir|S|JV0059 mreC protein - Escherichia coli >gp|M22055|ECOMREB_3 E.col	51	39
34	2	623	1327	gp|X16335|	Klebsiella pneumoniae rpoN gene 3'downstream region. [Klebsiella pneumoniae] >pir|S|S07661 hypothetical protein 162 (rpoN 3' region) - Klebsiella pneumoniae	51	32
37	11	6299	4824	pir|S|A2294 0	keratin, 67K type II cytoskeletal - human	51	40
39	1	188	3	gp|S78086|	protein-tyrosine phosphatase [Homo sapiens] >pir|S|B44929 protein-tyrosine-phosphatase (EC 3.1.3.48) BPTP-2 - human (fragment)	51	43
39	5	2837	1236	gp|Z22177|	F54G8.4 [Caenorhabditis elegans]	51	35
42	5	2265	3104	gp|M84415|	DNA polymerase [Bacteriophage SPO1] >pir|S|JC1269 DNA-directed DNA polymerase (EC 2.7.7.7) - phage SPO1	51	30
55	8	6786	7085	gp|L03292|	taurine/beta-alanine transporter [Mus cookii] >pir|S|A47194 taurine and beta-alanine transporter, TAUT - mouse	51	37
56	7	3620	3847	gp|K03277|	Tm2 gene product [Drosophila melanogaster] >pir|S|A25624 tropomyosin I, embryonic - fruit fly (Drosophila melanogaster)	51	25
66	4	3219	1651	gp|X78998|	endosomal protein [Homo sapiens] >pir|S|S44243 endosomal protein - human	51	22
75	6	2330	3406	gp|X73124|	ipa-83d gene product [Bacillus subtilis] >pir|S|S39738 hypothetical protein - Bacillus subtilis	51	34
89	4	1689	2285	gp|U32702|	DNA helicase [Haemophilus influenzae]	51	39
89	12	11391	9544	gp|L46170|	lic-1 operon protein [Haemophilus influenzae] >gp|U32829|HIU32829_8 lic-1 operon protein [Haemophilus influenzae] >gp|U00085|HIU00085_10 lic-1 operon protein	51	31

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					[Haemophilus influenzae] >gp|U32775|HIU32775_8 nucleotidyltransferase [Haemophilus influenzae] >p		
91	4	1983	1642	gp|L15626|	dynein [Saccharomyces cerevisiae]	51	25
98	1	754	2	gp|D64006|	hypothetical protein [Synechocystis sp.]	51	27
101	2	416	1012	gp|L45065|	hypothetical protein (SP:P33635) [Haemophilus influenzae] >gp|U32726|HIU32726_5 hypothetical protein (SP:P33635) [Haemophilus influenzae] >gp|U00073|HIU00073_46 hypothetical protein (SP:P33635) [Haemophilus influenzae] >gp|U32834|HIU32834_10 rRNA methylas	51	30
101	17	8473	8165	gp|L39923|	30S ribosomal protein S6 [Mycobacterium leprae]	51	30
101	35	21692	22174	gp|U10927|	CapD [Staphylococcus aureus]	51	28
102	3	2448	1180	gp|D64000|	hypothetical protein [Synechocystis sp.]	51	38
117	4	2669	2142	gp|L45352|	trigger factor [Haemophilus influenzae] >gp|U32754|HIU32754_4 trigger factor [Haemophilus influenzae] >gp|U00076|HIU00076_52 trigger factor [Haemophilus influenzae] >gp|U32700|HIU32700_4 peptidyl-prolyl cis-trans isomerase [Haemophilus influenzae] >pir|S|	51	27
128	2	2982	1534	gp|L44753|	hypothetical protein (GB:U14003_130) [Haemophilus influenzae] >gp|U32696|HIU32696_3 hypothetical protein (GB:U14003_130) [Haemophilus influenzae] >gp|U00070|HIU00070_14 hypothetical protein (GB:U14003_130) [Haemophilus influenzae] >gp|U32805|HIU32805_3 in	51	26
139	2	851	288	gp|D63999|	hypothetical protein [Synechocystis sp.]	51	35
572	1	519	235	gp|L15202|	Bacillus subtilis comE operon encoding ORF1, ORF2, ORF3 and Reverse-ORF genes, complete cds. [Bacillus subtilis] >pir|S|S39864 ComE ORF2 - Bacillus subtilis	51	33
8	3	3906	2470	gp|U23181|	ZK84.1 gene product [Caenorhabditis elegans]	50	28
11	11	5426	5923	gp|L45290|	rep helicase, single-stranded DNA-dependent ATPase [Haemophilus influenzae] >gp|U32748|HIU32748_3 rep helicase, single-stranded DNA-dependent ATPase [Haemophilus influenzae] >pir|S|D64084 rep helicase, single-stranded DNA-dependent ATPase (rep) homolog -	50	32

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

11	18	8259	9131	gp|U29399|	major outer sheath protein [Treponema denticola]	50	45
17	5	2963	2391	gp|U00008|	yejD [Escherichia coli]	50	33
22	14	9154	9777	gp|X59399|	AdgA protein [Rhodobacter capsulatus] >pir|S|S15555 adgA protein - Rhodobacter capsulatus	50	35
22	22	14326	14496	gp|X02307|	E. coli aspA gene for aspartase (L-aspartate ammonia-lyase) (EC 4.3.1.1). [Escherichia coli]	50	37
24	1	1	705	gp|M16152|	growth factor [Drosophila melanogaster]	50	33
27	8	4573	3698	gp|L46290|	ribonucleoside diphosphate reductase B2 subunit [Haemophilus influenzae] >gp|U32839|HIU32839_2 ribonucleoside diphosphate reductase B2 subunit [Haemophilus influenzae] >gp|U00086|HIU00086_23 ribonucleoside diphosphate reductase B2 subunit [Haemophilus inf	50	37
39	10	4905	6335	gp|L13845|	act206 gene product [Rhizobium meliloti]	50	24
43	5	4605	3028	gp|Z46729|	unknown [Saccharomyces cerevisiae] >pir|S|S49802 hypothetical protein YM9958.03c - yeast (Saccharomyces cerevisiae)	50	31
45	18	15712	16710	gp|U00010|	transport protein (similarity to antibiotic transport protein actII-3 from S.coelicolor) [Mycobacterium leprae]	50	24
53	10	5460	5296	gp|X78057|	carleticulin [Zea mays]	50	31
59	5	2553	2170	gp|U28375|	single-stranded DNA-specific exonuclease [Escherichia coli]	50	28
60	1	1	189	gp|U23764|	TonB [Pseudomonas aeruginosa]	50	32
61	6	1735	2823	gp|L44751|	hypothetical protein (GB:L01112_7) [Haemophilus influenzae] >gp|U32696|HIU32696_1 hypothetical protein (GB:L01112_7) [Haemophilus influenzae] >gp|U00070|HIU00070_12 hypothetical protein (GB:L01112_7) [Haemophilus influenzae] >gp|U32805|HIU32805_1 H. influ	50	32
62	2	954	403	gp|D64001|	hypothetical protein [Synechocystis sp.]	50	37
63	9	3455	3814	pir|S|S53457	dominant autoantigen gp 330 - rat (fragment)	50	50
70	32	19744	20682	gp|D64006|	hypothetical protein [Synechocystis sp.]	50	31
71	1	555	97	gp|X92441|	S.cerevisiae 33kb fragment from the right arm of chromosome XV. [Saccharomyces cerevisiae]	50	27
72	4	4363	2789	gp|X66793|	sigma factor 54 [Alcaligenes eutrophus] >pir|S|B48362 transcription initiation factor sigma 54 - Alcaligenes eutrophus	50	29

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

89	10	9208	8306	gp|X52905|	Escherichia coli betT, betI, betB and betA genes. [Escherichia coli] >pir|S|S15179 betT protein - Escherichia coli	50	37
96	2	173	658	gp|L45645|	hypothetical protein (GB:D10483_22) [Haemophilus influenzae] >gp|U32781|HIU32781_10 hypothetical protein (GB:D10483_22) [Haemophilus influenzae] >pir|S|A64164 hypothetical protein HI1007 - Haemophilus influenzae (strain Rd KW20) >gp|U00079|HIU00079_67 hyp	50	32
100	9	1781	2587	gp|X60395|	Hox-4.6 gene product [Mus musculus] >gp|X71422|MMHOXD11_1 HOXD-11 [Mus musculus] {SUB 14-336}	50	47
135	4	1991	2278	pir|S|S29717	adenylate cyclase (EC 4.6.1.1) type 5 - rat	50	34
143	5	5051	3876	gp|X94607|	phenylalanyl-tRNA synthetase alpha subunit [Saccharomyces cerevisiae]	50	27
581	1	20	454	gp|X69507|	endopeptidase I gene product [Bacillus sphaericus] >pir|S|S33310 endopeptidase I - Bacillus sphaericus >gp|X69895|BSPROTXA_1 peptidase I gene product [Bacillus sphaericus] {SUB 375-396}	50	40
639	1	1	489	gp|L44753|	hypothetical protein (GB:U14003_130) [Haemophilus influenzae] >gp|U32696|HIU32696_3 hypothetical protein (GB:U14003_130) [Haemophilus influenzae] >gp|U00070|HIU00070_14 hypothetical protein (GB:U14003_130) [Haemophilus influenzae] >gp|U32805|HIU32805_3 in	50	26
2	1	1	708	gp|X75627|	C.burnetii trxB, spoIIIE and serS genes. [Coxiella burnetii] >pir|S|S43133 hypothetical protein - Coxiella burnetii >pir|S|S31760 hypothetical protein Y - Coxiella burnetii	49	29
2	10	7878	6370	gp|L04520|	folyl-polyglutamate synthetase [Bacillus subtilis] >pir|S|B40646 folC - Bacillus subtilis	49	32
21	3	1389	2453	gp|U14003|	lysyl-tRNA synthetase analog [Escherichia coli] >pir|S|S56383 lysyl-tRNA synthetase genX - Escherichia coli >gp|X59988|ECGENXLTR_1 genX gene product [Escherichia coli] {SUB 11-335}	49	27
29	8	3666	3211	gp|U00013|	nifU [Mycobacterium leprae]	49	30
30	15	10668	12164	gp|L44676|	penicillin-binding protein 2 [Haemophilus influenzae]	49	34

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					>gp|U32688|HIU32688_11 penicillin-binding protein 2 [Haemophilus influenzae] >gp|U00069|HIU00069_30 penicillin-binding protein 2 [Haemophilus influenzae] >gp|U32797|HIU32797_13 penicillin-binding prote		
43	1	1	1224	gp|M13359|	E.coli sppA gene encoding protease IV, complete cds. [Escherichia coli] >pir|S|PRECT4 proteinase IV (EC 3.4.-.-) - Escherichia coli >gp|U13772|ECU13772_1 protease IV [Escherichia coli] {SUB 110-433}	49	37
44	13	7058	7858	gp|L45206|	hypothetical protein (SP:P32662) [Haemophilus influenzae] >gp|U32738|HIU32738_4 hypothetical protein (SP:P32662) [Haemophilus influenzae] >gp|U00074|HIU00074_85 hypothetical protein (SP:P32662) [Haemophilus influenzae] >gp|U32847|HIU32847_4 hydrolase (pho	49	31
47	14	8255	7647	gp|M92905|	calcium channel alpha-1 subunit [Rattus norvegicus]	49	36
60	3	1212	586	gp|D17462|	Na+ -ATPase subunit D [Enterococcus hirae] >pir|S|F53610 ntpD protein - Enterococcus hirae	49	25
79	17	8989	10035	gp|L14556|	glutamate synthase [Escherichia coli] >gp|U00006|ECOUW89_26 glutamate synthase [Escherichia coli] >gp|V00347|ECRRNB_1 E. coli 3' noncoding region preceding the gene rrnB which codes for 16S ribosomal RNA. [Escherichia coli] >gp|V00348|ECRRNBZ_1 E. coli ri	49	34
15	3	855	1724	gp|J03998|	P.falciparum glutamic acid-rich protein gnen, complete cds. [Plasmodium falciparum] >pir|S|A54514 glutamic acid-rich protein precursor - Plasmodium falciparum	48	21
16	6	4359	4063	gp|X65195|	N-acetylphosphinothricin-tripetide- deacetylase [Streptomyces viridochromogenes] >pir|S|S20686 N-acetylphosphinothricin-tripetide-deacetylase - Streptomyces viridochromogenes >gp|M22827|STMPAT_3 Streptomyces viridochromogenes phosphinothricin N-acetyltran	48	35
16	39	23653	22703	gp|U21320|	K04G7.3 gene product [Caenorhabditis elegans]	48	27
26	36	23309	24658	gp|U13961|	protective surface antigen D15 [Haemophilus influenzae] >pir|S|JC4078 D-15 protective surface antigen - Haemophilus influenzae	48	29

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

29	25	11464	12030	gp|U20217|	fibrillin-2 [Mus musculus]	48	36
32	3	1511	3190	gp|L44947|	apolipoprotein N-acyltransferase [Haemophilus influenzae] >gp|U32716|HIU32716_3 apolipoprotein N-acyltransferase [Haemophilus influenzae] >gp|U00072|HIU00072_32 apolipoprotein N-acyltransferase [Haemophilus influenzae] >gp|U32825|HIU32825_3 apolipoprotein	48	30
41	7	1910	2155	gp|L36829|	alphaA-crystallin-binding protein I [Mus musculus] >gp|X68946|MMACRYBP1_1 alphaA-CRYBP1 [Mus musculus] {SUB 2024-2688}	48	45
42	16	9061	9687	gp|D50303|	Ribosomal Protein L10 [Bacillus subtilis]	48	29
47	23	13542	12838	gp|L44848|	hypothetical protein (SP:P21504) [Haemophilus influenzae] >gp|U32705|HIU32705_7 hypothetical protein (SP:P21504) [Haemophilus influenzae] >gp|U00071|HIU00071_18 hypothetical protein (SP:P21504) [Haemophilus influenzae] >gp|U32814|HIU32814_7 H. influenzae	48	30
49	10	7561	7097	gp|D26185|	expressed at the end of exponential growyh under condtions in which the enzymes of the TCA cycle are repressed [Bacillus subtilis] >gp|X16518|BSTMSPRS_3 B.subtilis prs,tms,and ctc (partial) genes for PRPP synthetase and two undefined gene products. [Bacil	48	26
63	6	2557	1895	gp|Z33126|	membrane forming protein [Mycoplasma capricolum] >pir|S|S48611 hypothetical protein - Mycoplasma capricolum (SGC3) (fragment) {SUB 1-101}	48	26
85	2	336	764	gp|L33468|	thermoregulated motility protein [Yersinia enterocolitica type 0:8] >gp|L33468|YEPTMA_1 thermoregulated motility protein [Yersinia enterocolitica (type 0:8)]	48	24
100	20	5688	4987	gp|S72442|	peptidyl-prolyl cis/trans isomerase [Legionella pneumophila] >pir|S|A30591 outer membrane protein mip precursor - Legionella pneumophila	48	31
101	5	1578	1841	gp|M96434|	oxaloacetate decarboxylase [Salmonella typhimurium]	48	35
117	9	6598	6206	gp|L40632|	ankyrin 3 [Mus musculus]	48	30
564	1	2	385	gp|L44946|	hemolysin [Haemophilus influenzae] >gp|U32716|HIU32716_2	48	25

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					hemolysin [Haemophilus influenzae] >gp|U00072|HIU00072_31 hemolysin [Haemophilus influenzae] >gp|U32825|HIU32825_2 cystathionine beta-synthase [Haemophilus influenzae] >pir|S|G64060 hemolysin (tlyC)		
1	26	11879	12952	gp|L46316|	nitrogen fixation protein [Haemophilus influenzae] >gp|U32841|HIU32841_8 nitrogen fixation protein [Haemophilus influenzae] >gp|U00086|HIU00086_48 nitrogen fixation protein [Haemophilus influenzae] >gp|U32788|HIU32788_4 membrane protein, NADH:ubiquinone o	47	27
8	16	8098	8844	gp|L45104|	oxygen-independent coproporphyrinogen III oxidase [Haemophilus influenzae] >gp|U32729|HIU32729_6 oxygen-independent coproporphyrinogen III oxidase [Haemophilus influenzae] >gp|U00073|HIU00073_85 oxygen-independent coproporphyrinogen III oxidase [Haemophil	47	29
16	37	22283	22083	gp|Z33413|	unknown [Pseudomonas syringae] >gp|Z33413|PSFOSCG_3 unknown [Pseudomonas syringae] >pir|S|S44937 hypothetical protein - Pseudomonas syringae	47	23
27	7	3449	3781	gp|Z49782|	ywkE gene product [Bacillus subtilis] >pir|S|S55438 ywkE protein - Bacillus subtilis	47	36
29	13	5615	4911	gp|U00013|	pps1 [Mycobacterium leprae]	47	23
37	14	7473	8843	gp|L47709|	ypiA gene product [Bacillus subtilis]	47	24
55	11	10980	9094	gp|U33007|	D9461.18p [Saccharomyces cerevisiae]	47	30
57	13	9939	10847	gp|X64259|	N-acetylglucosaminyl transferase [Bacillus subtilis] >pir|S|JC1275 phospho-N-acetylmuramoyl-pentapeptide-transferase (EC 2.7.8.13) - Bacillus subtilis >gp|M31827|BACDDSA_1 Bacillus subtilis (clone lambda-BS1) cell division and sporulation protein (dds) ge	47	27
63	8	4537	3023	gp|M13169|	high affinity ribose transport protein [Escherichia coli] >pir|S|B26304 ribose transport protein rbsA - Escherichia coli	47	27
104	18	7200	8012	gp|U00021|	Mycobacterium leprae cosmid L247. [Mycobacterium leprae]	47	30
158	1	1331	15	gp|U21734|	UNC-44 [Caenorhabditis elegans] >pir|S|A57282 ankyrin-related protein unc-44 - Caenorhabditis elegans (fragment)	47	34

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

175	3	1415	729	gp|U34923|	ThiJ [Escherichia coli]	47	26
326	1	322	2	gp|M83196|	microtubule-associated protein 1A [Rattus norvegicus] >pir|S|A43359 microtubule-associated protein MAP1A - rat	47	38
36	22	17495	18919	gp|X77395|	uridine kinase [Saccharomyces cerevisiae] >gp|X53998|SCURK1_2 uridine kinase [Saccharomyces cerevisiae] >pir|S|S29374 uridine kinase (EC 2.7.1.48) - yeast (Saccharomyces cerevisiae)	46	27
37	7	2757	3395	gp|D30808|	Glutathione-regulated potassium efflux system (KcfC) [Bacillus subtilis]	46	30
135	3	3028	878	gp|L45970|	hypothetical protein (GB:L12968_1) [Haemophilus influenzae] >gp|L46094|HEAHI1463_1 hypothetical protein (GB:L12968_1) [Haemophilus influenzae] >gp|U32813|HIU32813_2 hypothetical protein (GB:L12968_1) [Haemophilus influenzae] >gp|U32824|HIU32824_5 hypothet	46	34
1	15	6998	7420	gp|M16489|	Escherichia coli tolQRA gene cluster DNA. [Escherichia coli] >pir|S|WMEC15 15.5K protein (tolAB operon 5' region) - Escherichia coli >gp|U30934|ECU30934_2 Escherichia coli cytochrome oxidase d subunit II (cydB) gene, partial cds, and orf in tolQRA region,	45	30
10	8	5476	5099	gp|Z48239|	orf7 gene product [Saccharomyces cerevisiae] >pir|S|S57679 hypothetical protein 7 - yeast (Saccharomyces cerevisiae)	45	24
12	6	4020	4358	gp|U34774|	ankyrin-like repeat protein [Orf virus] >gp|S78516|S78516_1 G1L gene product [Unknown.]	45	36
17	14	7538	7765	gp|Z48236|	X-prolyl dipeptidyl aminopeptidase [Lactobacillus helveticus]	45	27
22	16	10438	11088	gp|L46259|	dedA protein [Haemophilus influenzae] >gp|U32836|HIU32836_3 dedA protein [Haemophilus influenzae] >gp|U00085|HIU00085_96 dedA protein [Haemophilus influenzae] >gp|U32782|HIU32782_16 alkaline phosphatase-like protein [Haemophilus influenzae] >pir|S|D64133	45	27
44	12	5893	7155	gp|L44883|	H. influenzae predicted coding region HI0238 [Haemophilus influenzae] >gp|U32710|HIU32710_4 H. influenzae predicted coding region HI0238 [Haemophilus influenzae] >gp|U00071|HIU00071_54 H. influenzae predicted coding region	45	22

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					HI0238 [Haemophilus influenzae]		
81	17	14515	16290	gp|Z32522|	ligoendopeptidase F [Lactococcus lactis] >pir|S|S49150 ligoendopeptidase F - Lactococcus lactis	45	28
100	1	1	1059	gp|Z66499|	T01B7.8 [Caenorhabditis elegans]	45	39
104	17	6278	7399	gp|X73124|	ipa-65d gene product [Bacillus subtilis] >pir|S|S39720 hypothetical protein - Bacillus subtilis	45	24
13	10	6339	8366	gp|U13675|	lactose permease [Citrobacter freundii] >pir|S|JC2544 lactose carrier protein - Citrobacter freundii	44	31
26	35	22077	23420	gp|U13961|	protective surface antigen D15 [Haemophilus influenzae] >pir|S|JC4078 D-15 protective surface antigen - Haemophilus influenzae	44	29
35	10	6451	7464	gp|D90109|	Rat mRNA for long-chain acyl-CoA synthetase (EC 6.2.1.3). [Rattus norvegicus] >gp|M55642|RATCOAA_1 acyl-CoA synthetase [Rattus norvegicus] >pir|S|A36275 long-chain-fatty-acid--CoA ligase (EC 6.2.1.3) - rat	44	30
41	1	23	751	gp|M21994|	E.coli cysK gene, 3' end, ptsH, ptsI, and crr phototransferase system genes, complete cds. [Escherichia coli] >gp|J02796|ECOPTSHI_2 ptsI gene product [Escherichia coli] >pir|S|WQECPI phosphotransferase system enzyme I (EC 2.7.3.9) - Escherichia coli	44	27
47	13	7765	6896	gp|S44426|	cytB gene product [Synechococcus] >pir|S|PQ0180 CytB protein - Synechococcus sp. (PCC 7942) (fragment)	44	32
98	5	4080	4325	gp|Z11582|	nuf1 gene product [Saccharomyces cerevisiae] >gp|X73297|SCSETRP4_2 SPC110/NUF1 gene product [Saccharomyces cerevisiae] >gp|U28372|YSCD9476_7 Probable essential component of the nucleoskeleton (Swiss Prot. accession number P32380) [Saccharomyces cerevisiae	44	24
117	7	4540	5535	gp|Z54142|	unknown [Schizosaccharomyces pombe] >gp|X92894|SPHXK1P_1 hexokinase 1 [Schizosaccharomyces pombe]	44	27
152	3	2944	2366	gp|M35200|	M.xanthus frzG and frzF genes, complete cds. [Myxococcus xanthus] >pir|S|XYYZFG frzG protein - Myxococcus xanthus	44	29

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

153	1	748	2	gp|X82489|	pyruvate,orthophosphate dikinase [Mesembryanthemum crystallinum] >gp|X82489|MCPPD_1 pyruvate,orthophosphate dikinase [Mesembryanthemum crystallinum] >pir|S|S55478 pyruvate,orthophosphate dikinase (EC 2.7.9.1) - common ice plant	44	36
156	1	2	340	gp|L14321|	BICP4 [Bovine herpesvirus type 1]	44	38
183	3	806	1117	gp|U02514|	hook-associated protein 3 [Escherichia coli] >pir|S|S44022 hook-associated protein 3 - Escherichia coli	44	30
1	25	11110	11946	gp|X72888|	rnfC gene product [Rhodobacter capsulatus] >pir|S|S39893 rnfC protein - Rhodobacter capsulatus >gp|X79064|RCRNFCD_1 rnfC gene product [Rhodobacter capsulatus] {SUB 39-519}	43	28
11	12	5921	6973	gp|M63489|	ATP-dependent nuclease [Bacillus subtilis] >pir|S|B39432 ATP-dependent exonuclease synthesis protein AddA - Bacillus subtilis	43	26
42	14	8004	7747	gp|U24265|	special lobe-specific protein [Chironomus thummi]	43	33
64	30	12557	13519	gp|L45532|	membrane fusion protein [Haemophilus influenzae] >gp|U32771|HIU32771_6 membrane fusion protein [Haemophilus influenzae] >gp|U32717|HIU32717_6 permease [Haemophilus influenzae] >pir|S|G64100 membrane fusion protein (mtrC) homolog - Haemophilus influenzae (	43	22
77	8	3699	4703	pir|S|S02708	troponin T - fruit fly (Drosophila melanogaster)	43	29
85	9	3407	3718	gp|X64346|	Herpesvirus saimiri complete genome DNA. [Saimirine herpesvirus 1] >pir|S|A36811 hypothetical protein ORF48 - saimiriine herpesvirus 1 (strain 11) >gp|M86409|HSV3PRGEN_1 Herpesvirus saimiri the most three prime end of the genome. [Herpesvirus saimiri] {SU	43	35
159	1	2	1165	gp|L35574|	sigma-B regulator [Bacillus subtilis]	43	20
162	1	302	1714	gp|L23195|	cytoplasmic dynein heavy chian [Drosophila melanogaster] >pir|S|A54794 dynein heavy chain, cytoplasmic - fruit fly (Drosophila melanogaster) >gp|L25122|DRODYNEINH_1 dynein heavy chain [Drosophila melanogaster] {SUB 1877-1998}	43	26
1	13	5428	6120	gp|Z67757|	unknown [Schizosaccharomyces pombe]	42	31
6	5	4911	3877	gp|X77091|	trkA gene product [Escherichia coli] >gp|U18997|ECOUW67_214	42	28

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

					TrkA protein of the constitutive K+ transport system Trk [Escherichia coli] >gp|X52114|ECTRKAG_3 TrkA protein of the constitutive K+ -transport system Trk [Escherichia coli] >pir|S|S36252 trkA pr		
16	5	2829	4070	gp|L42115|	insulin-activated amino acid transporter [Mus musculus] >pir|S|JC4149 adipocyte amino acid transporter - mouse	42	24
100	4	1738	1061	gp|X88798|	AT1 gene product [Oryza sativa] >pir|S|S57459 hook-containing protein AT1 - rice	42	37
29	14	5713	5516	gp|U03994|	envelope protein [Simian immunodeficiency virus] >pir|S|S46335 envelope protein - simian immunodeficiency virus	41	27
53	8	3943	3542	gp|D10474|	ORF248 [Synechocystis sp.] >pir|S|JT0603 hypothetical 27.8K protein (frxC 3' region) - Synechocystis sp. (PCC 6803)	41	26
65	3	356	595	gp|A15841|	proteinase gene product [Lactococcus lactis cremoris]	41	29
68	2	1774	572	gp|X06545|	E. coli genes hsdR and hsdM. [Escherichia coli] >pir|S|NDECKR type I site-specific deoxyribonuclease (EC 3.1.21.3) EcoK chain R - Escherichia coli	41	24
70	30	18836	20197	gp|U00001|	CDC27 [Homo sapiens]	41	23
101	20	10060	11289	gp|L31959|	Mus musculus (strain C3HF/RL) ORF mRNA, complete cds. [Mus musculus]	41	26
2	6	3975	4871	gp|D26185|	high level kasgamycin resistance [Bacillus subtilis]	40	27
16	1	993	4	gp|U41010|	T05A12.2 gene product [Caenorhabditis elegans]	40	27
47	7	4032	2482	gp|X06165|	Yeast CDC16 gene. [Saccharomyces cerevisiae] >gp|Z28022|SCYKL022C_1 CDC16 gene product [Saccharomyces cerevisiae] >pir|S|A27832 cell division control protein CDC16 - yeast (Saccharomyces cerevisiae)	40	27
11	17	7969	7592	gp|Y00063|	Plasmodium falciparum mRNA fragment for knob protein. [Plasmodium falciparum] >pir|S|S14431 Knob-associated histidine-rich protein - Plasmodium falciparum (fragment)	39	25
234	1	3	302	gp|M87634|	BF-1 [Rattus norvegicus] >pir|S|JH0672 brain factor 1 protein - rat	39	32
56	12	5977	6690	gp|X65165|	extensin [Volvox carteri] >pir|S|S22697 extensin - Volvox carteri (fragment)	38	33
70	5	2560	4332	gp|D43965|	Xanthine Dehydrogenase [Bombyx mori]	36	21

TABLE 2.

Treponema pallidum - Putative coding regions of novel proteins similar to know proteins

56	13	6826	6131	gp|U22680|	X box repressor [Homo sapiens]	35	19
37	13	5734	7578	gp|Z38061|	mal5 gene product [Saccharomyces cerevisiae] >gp|Z47047|SCCHRIX_196 Sta1p [Saccharomyces cerevisiae] >pir|S|S48478 glucan 1,4-alpha-glucosidase (EC 3.2.1.3) homolog - yeast (Saccharomyces cerevisiae) >gp|M16165|YSCS22_1 Yeast (S.cerevisiae) S1 protein gen	33	19

TABLE 3.

Treponema pallidum - Putative coding regions of novel proteins not similar to know proteins

Contig ID	ORF ID	Start (nt)	Stop (nt)
1	8	2998	2366
1	9	2391	2651
1	14	6072	7019
1	19	8764	8958
1	20	9778	10317
1	21	9930	10325
1	22	10323	11285
1	23	10832	10473
1	24	11294	10878
1	27	13370	13990
2	2	702	1268
2	3	1027	1950
2	4	2199	1417
2	7	4886	5182
2	8	5095	5343
2	16	13354	13947
4	1	1360	470
4	6	4285	4848
4	10	7150	6740
4	14	9207	8596
5	8	5721	5383
6	1	442	2
6	6	4237	4839
7	1	29	433
8	1	2321	1545
8	2	2491	2105
8	4	2761	3090
8	5	3326	3568
8	6	3566	3751
8	7	4421	3717
8	8	4388	4885
8	9	4785	5204
8	11	6460	5501
8	13	6741	7334
8	14	7189	7557
8	18	9207	8998
8	38	23258	23581
8	39	23766	23395
9	1	1	504
9	2	2312	588
9	3	3895	2333
9	4	3116	3475
10	2	2252	2515
10	3	3473	2904
10	4	4003	3419
10	5	4305	3829

TABLE 3.

Treponema pallidum - Putative coding regions of novel proteins not similar to know proteins

10	6	4439	4107
10	10	6172	6744
10	11	6546	7367
10	12	7175	7546
10	13	7792	7595
10	14	8345	7830
10	15	8248	7940
10	16	8292	9110
10	17	9101	8778
11	1	122	469
11	2	849	1358
11	3	1129	1563
11	4	1624	2919
11	5	2826	3164
11	6	3047	3991
11	7	3940	4557
11	9	5188	4898
11	10	5190	5477
11	13	6392	6111
11	16	8220	7573
11	20	9910	10440
11	21	10917	10402
11	22	10438	11469
11	23	11607	11053
11	24	11950	11576
11	26	13738	14235
11	27	14115	14321
12	1	40	585
12	2	2403	352
12	5	3686	3952
12	9	4715	4957
13	2	680	1303
13	3	1613	747
13	5	2555	1920
13	7	3468	3130
13	9	6157	5279
13	13	8327	8638
15	1	1	759
16	9	6239	5763
16	11	7440	7069
16	12	7901	7194
16	15	9565	10185
16	17	12182	11706
16	18	12584	12300
16	21	12793	13308
16	22	14166	13264
16	23	14450	13992

TABLE 3.

Treponema pallidum - Putative coding regions of novel proteins not similar to know proteins

16	26	16277	15594
16	31	17895	18332
16	32	19003	18242
16	33	19491	18970
16	34	20385	20053
16	40	22895	23161
16	42	24264	24581
16	43	24452	25549
16	45	26020	26673
16	46	26901	27722
16	53	31056	31415
16	55	33441	32989
17	2	1345	962
17	3	2034	1492
17	7	3509	3105
17	8	4602	3892
17	10	5335	4946
17	11	5611	5381
17	12	5645	6409
17	13	6744	7742
17	15	8000	7836
17	16	7839	8450
17	17	8619	8434
18	6	5410	5991
18	8	6421	6020
19	6	4526	4951
19	8	6964	7773
19	9	7544	8032
19	10	8049	8486
19	12	9692	10201
19	15	12883	13215
19	17	13106	13624
19	23	18224	17556
19	24	19216	18200
20	2	1814	2575
20	3	2511	3299
21	13	9403	9813
21	14	9489	9917
21	17	10312	10758
21	19	10602	10946
22	2	1671	2450
22	3	3504	2578
22	4	3372	3869
22	6	5309	4761
22	7	6062	5445
22	8	5592	5975
22	9	6144	7343

TABLE 3.

Treponema pallidum - Putative coding regions of novel proteins not similar to know proteins

22	10	6416	6790
22	11	8006	7293
22	12	7850	8323
22	13	8617	9156
22	18	12328	13269
23	4	1517	1212
24	3	811	1548
25	3	1729	1169
26	2	167	1576
26	3	1512	2843
26	8	4822	4983
26	9	5604	4888
26	10	5855	6100
26	11	6040	6558
26	13	7758	7267
26	14	8349	7903
26	18	9960	10109
26	30	17705	19171
26	31	18835	19236
26	32	19161	19769
26	33	19688	22006
26	34	20812	20591
27	1	669	4
27	2	788	195
28	5	3158	2616
28	6	3367	3071
28	9	4459	4707
28	12	5186	5794
28	15	6636	7073
28	16	7839	7339
28	17	7971	8477
28	18	8389	9942
28	20	9831	10127
28	21	9914	10801
29	3	804	1616
29	4	1519	2061
29	5	1976	2404
29	9	3478	3705
29	12	4800	5204
29	15	6102	5584
29	18	7459	7866
29	21	9443	8706
29	27	12452	13120
30	1	111	1217
30	3	1735	1517
30	16	12128	12544
30	18	13844	14488

TABLE 3.

Treponema pallidum - Putative coding regions of novel proteins not similar to know proteins

34	5	3334	3017
34	6	3095	4246
35	2	856	230
35	3	1968	514
35	4	2807	2073
35	5	3424	2282
35	9	6057	6809
35	11	6867	6592
35	12	7768	7217
35	13	8548	7856
35	17	10143	9238
35	18	9728	10747
35	19	11411	10851
35	35	19000	18431
35	39	21598	21347
36	4	3846	3448
36	5	4694	3759
36	6	4920	5420
36	7	4991	5542
36	8	5540	5842
36	9	6906	5854
36	10	5950	6246
36	11	6816	7121
36	12	7418	6858
36	23	18768	19202
36	25	20011	19688
37	2	1221	751
37	4	1815	1438
37	5	2429	1971
37	9	4315	3716
37	12	5504	6412
37	15	9075	9386
37	17	11158	10835
38	4	1964	1626
38	5	1791	2450
39	2	826	134
39	4	1353	781
39	6	2660	3136
39	7	3336	2962
39	11	6254	5331
40	4	1580	2065
40	5	1807	2034
41	6	2078	1716
41	10	3059	3976
41	11	3865	4854
41	12	5143	4844
41	15	5944	6429

TABLE 3.

Treponema pallidum - Putative coding regions of novel proteins not similar to know proteins

42	2	931	1203
42	4	2149	1514
42	25	18100	17825
43	2	1542	1931
43	4	2842	3246
44	1	351	67
44	3	534	349
44	4	825	532
44	10	5146	5505
44	11	5788	5171
44	15	9156	8251
44	20	12138	12524
45	2	2281	1058
45	8	8147	8656
45	9	8437	8760
45	11	9062	9373
45	15	13884	12790
45	17	14500	15108
47	3	1317	1550
47	5	2545	2105
47	6	2406	3014
47	9	4837	4592
47	10	5173	5700
47	15	8714	9049
47	17	10561	10001
47	19	11772	10807
47	24	13781	13425
47	27	15122	14625
47	28	15917	15480
47	30	16213	16638
48	2	1210	725
48	3	2133	1198
48	5	3031	2483
48	13	9988	9617
48	14	10323	9721
48	15	10538	10149
48	16	11509	10493
48	17	11818	11507
48	20	12934	12656
49	2	1116	823
49	3	866	1435
49	4	1667	2584
49	5	3915	3295
49	6	5194	3806
49	7	5691	5113
49	14	9465	9136
49	15	9467	9760

TABLE 3.

Treponema pallidum - Putative coding regions of novel proteins not similar to know proteins

49	16	10114	9578
50	2	252	611
51	2	3123	3671
51	7	7326	7664
52	1	2445	1756
52	2	2956	2459
52	4	4723	4109
53	1	3	365
53	2	172	603
53	4	1190	924
53	5	1755	1411
53	6	3185	1647
54	16	5423	6022
54	17	5994	6335
54	21	7481	7143
54	22	7243	8028
54	23	7289	7525
54	24	7668	7985
55	2	703	464
55	6	4738	6135
55	7	5937	6731
55	10	8692	9015
56	10	6010	5567
56	11	5699	6061
57	1	3	446
57	10	7281	7565
58	2	1204	551
58	5	2883	2335
59	2	568	867
59	3	2094	1591
59	4	2239	1871
60	2	637	2
60	7	5121	4279
61	2	666	1283
61	3	1263	976
61	4	1281	1778
61	5	1736	1392
61	7	2988	2608
61	10	4819	5673
62	5	3223	2078
62	6	2669	2959
62	7	3389	3099
63	2	553	341
63	3	1189	893
63	4	1916	966
63	5	1729	1992
63	7	3177	2464

TABLE 3.

Treponema pallidum - Putative coding regions of novel proteins not similar to know proteins

63	10	3552	3887
63	11	3926	4531
63	16	6897	5938
63	17	6446	6099
63	18	7228	6995
63	26	14887	14240
63	28	17256	16678
63	29	17509	17033
63	30	17855	17562
63	33	19853	19116
63	35	20299	21408
64	1	236	45
64	2	1029	196
64	6	2960	3196
64	7	3785	3159
64	15	6989	7399
64	16	7273	7719
64	17	7692	8126
64	18	8250	8816
64	19	8749	9192
64	20	9060	9332
64	22	9330	10403
64	23	10052	9453
64	24	10256	10879
64	26	10922	11665
64	27	11590	11970
64	29	12472	12233
65	1	43	378
65	6	2556	3329
65	7	3194	4339
65	8	4105	3431
65	9	4176	4448
66	3	1622	1425
68	3	2093	1686
68	4	2316	2119
68	5	2399	3049
68	6	3013	2450
68	8	3390	2881
69	1	3	743
69	2	629	886
69	3	1397	906
69	4	2091	1342
70	4	2237	2662
70	8	6703	6996
70	9	7148	7954
70	10	7671	8378
70	12	8925	9299

TABLE 3.

Treponema pallidum - Putative coding regions of novel proteins not similar to know proteins

70	13	9103	9369
70	15	10187	11269
70	16	11447	11022
70	17	12491	11793
70	19	13758	13964
70	24	16029	15769
70	27	17085	17675
70	29	18387	17896
70	31	19514	18957
70	33	20556	20128
71	3	1350	1105
72	5	3863	4087
74	1	129	611
74	2	676	380
74	3	440	916
74	4	820	1728
74	6	2031	1285
74	7	1820	1545
74	9	3676	4929
74	10	4994	6424
74	12	7476	7748
74	13	7801	7634
74	14	7640	8152
75	1	748	2
75	2	1221	655
75	3	2151	1882
75	4	3249	2212
75	5	2233	2559
75	7	2932	2375
75	8	2683	3639
75	9	3507	4034
75	10	3777	3559
75	12	4514	5227
75	13	5377	4922
75	14	5861	5556
77	3	1365	2441
77	4	2317	2832
77	5	2514	2888
77	6	2816	3664
77	7	3517	3912
77	9	4475	4855
77	10	4788	5060
79	6	2455	2727
79	7	3038	2772
79	13	7889	7512
79	15	8631	8059
79	16	8802	9056

TABLE 3.

Treponema pallidum - Putative coding regions of novel proteins not similar to know proteins

79	18	10273	9752
79	19	10092	10436
80	3	1228	551
81	2	1024	611
81	3	1970	888
81	4	1605	1997
81	10	7615	8082
81	16	14381	14085
81	18	15195	14815
82	1	3	260
83	1	1	897
83	2	351	866
83	5	1759	2667
83	6	2565	2984
83	7	2932	3282
83	9	3362	3784
83	10	4123	3494
83	12	4452	4105
83	17	7929	9203
84	5	1484	1813
84	6	2019	1810
84	7	2829	2377
84	8	3531	4421
85	1	1	510
85	3	762	1235
86	4	2070	1549
86	5	1600	2073
87	3	2608	2021
87	4	2615	2274
88	2	894	2081
88	3	2625	1975
88	7	3403	3750
89	2	992	1432
92	2	660	1043
93	1	17	580
93	2	626	1348
94	3	1869	1480
94	4	1706	2074
95	1	2	184
95	2	37	327
95	4	1018	1494
95	5	1685	1284
95	6	1881	1624
96	3	586	191
98	4	4094	3825
98	6	4968	4618
100	2	1356	4

TABLE 3.

Treponema pallidum - Putative coding regions of novel proteins not similar to know proteins

100	3	653	183
100	5	1095	1532
100	6	1283	1585
100	7	1450	1836
100	8	2010	1459
100	10	2149	2619
100	13	3049	3573
100	16	3631	4023
100	23	7950	7630
100	24	7690	8220
100	25	8201	7890
100	26	8592	8990
100	27	10023	8770
100	35	14886	15329
100	36	15272	15769
100	38	16962	17621
100	40	18590	18327
101	3	802	515
101	18	8545	9162
101	22	13204	12545
101	24	13821	14135
101	27	16471	16788
101	28	17178	16543
101	29	18178	17063
101	30	18212	18847
101	32	19052	19798
102	2	1415	921
102	4	2083	2730
102	8	5114	4851
102	9	5420	5112
103	3	1699	2436
103	6	4533	5243
104	2	473	1069
104	3	1085	1711
104	4	1411	1695
104	6	1896	2405
104	8	3705	3145
104	9	3729	3983
104	10	4782	4192
104	14	5889	5557
104	16	6123	6404
105	1	472	68
105	2	890	375
105	3	849	1592
107	1	1575	898
107	2	2213	1503
107	3	2994	2173

TABLE 3.

Treponema pallidum - Putative coding regions of novel proteins not similar to know proteins

107	4	4979	2892
109	2	1429	497
110	2	698	1399
110	3	1396	1133
111	3	1854	1486
112	3	2233	1868
112	4	3078	1915
112	13	9340	9005
114	2	576	190
115	3	1551	1090
115	5	2499	1933
116	1	346	5
116	2	439	1143
116	3	801	478
116	4	1254	919
116	6	1646	1188
116	8	1968	2459
116	11	3583	4077
116	12	4235	4852
116	14	5140	6093
116	15	6922	6170
116	16	6483	6722
117	8	6315	5743
117	10	6984	6493
121	2	604	1314
121	9	3632	4156
121	10	4524	5723
121	11	6081	5785
121	12	6015	6359
121	13	6598	7308
121	15	7894	7637
121	17	8347	8027
121	20	9966	10946
121	21	10457	10110
121	28	16019	15696
121	29	17161	16733
121	30	16754	17380
121	31	18342	17401
122	1	325	729
122	2	603	1808
122	4	2300	2785
126	1	191	3
126	3	906	724
126	4	1484	2170
134	1	486	283
135	2	880	554
135	5	2797	3129

TABLE 3.

Treponema pallidum - Putative coding regions of novel proteins not similar to know proteins

135	8	3381	3695
137	1	2	970
137	2	180	22
137	4	1037	1321
137	5	1319	3001
137	6	1966	1601
137	7	2707	3690
137	8	3501	3169
137	9	3414	4613
137	10	4726	4535
137	12	5477	5160
139	3	819	1229
140	3	1465	671
141	2	1273	848
141	7	4071	4331
141	9	4761	5282
142	7	10511	9729
142	8	10514	11344
142	9	11325	12233
142	10	13391	12678
143	6	5906	5304
143	7	6546	6130
144	3	640	1065
144	5	1627	1121
144	6	2102	1593
144	7	2704	1979
144	8	2898	2344
145	1	3	983
145	2	728	51
145	3	763	341
145	4	890	2065
145	5	1908	1570
145	6	1927	2478
145	7	2499	2011
145	8	2304	2669
145	9	2629	2967
145	10	3141	2638
145	11	2877	3290
146	1	3	1316
148	1	1	261
148	2	1145	282
149	1	16	489
150	4	3743	3285
150	10	8189	8986
150	12	11486	10512
150	13	10524	10994
150	15	12079	12366

TABLE 3.

Treponema pallidum - Putative coding regions of novel proteins not similar to know proteins

152	2	2411	1977
156	3	561	139
156	4	883	635
157	1	1	762
157	2	662	1231
157	4	2981	1695
158	2	456	43
161	1	466	5
161	3	558	857
163	1	607	140
163	2	716	468
164	1	1283	3
165	2	369	88
169	2	343	564
172	2	609	295
172	5	1978	1466
172	6	2076	1792
172	7	2825	2019
172	8	2424	2864
178	1	98	556
182	1	248	3
186	3	734	1267
186	4	1113	1379
188	1	3	686
188	2	310	843
189	1	688	2
192	1	245	3
192	2	18	413
193	1	85	507
199	1	146	376
203	1	321	611
206	1	2	568
209	1	1	543
210	1	229	2
212	1	42	584
212	2	383	808
224	1	38	286
224	2	579	325
276	1	201	587
328	1	360	4
376	1	567	139
389	1	485	3
423	1	545	270
478	1	277	11
480	1	27	305
482	1	327	79
484	1	310	8

TABLE 3.

Treponema pallidum - Putative coding regions of novel proteins not similar to know proteins

522	1	473	75
524	1	248	3
551	1	8	547
558	1	3	455
559	1	537	55
565	1	82	420
566	2	360	929
566	3	769	1104
579	2	379	2
605	1	334	53
608	1	186	4
620	1	444	115
625	1	281	3
626	1	253	41
626	3	847	578
628	1	555	79
628	2	626	306
633	1	195	4
634	2	35	583
636	1	3	308
643	1	1	402
644	1	1	339
644	2	525	4
645	3	747	427
646	1	79	453
648	1	426	4
649	1	264	536
659	1	90	359
668	1	103	342
668	3	288	536
669	1	251	39
678	1	382	95
679	1	513	130
682	1	108	434
684	1	438	133
687	1	2	262
691	1	337	14
702	1	549	121
703	1	2	307
719	1	531	358
742	1	408	220

(1) GENERAL INFORMATION:

(i) APPLICANT: Human Genome Sciences Inc., et. al.

(ii) TITLE OF INVENTION: Treponema pallidum Polynucleotides and Sequences

(iii) NUMBER OF SEQUENCES: 744

(iv) CORRESPONDENCE ADDRESS:

(A) ADDRESSEE: Human Genome Sciences, Inc.

(B) STREET: 9410 Key West Avenue

(C) CITY: Rockville

(D) STATE: Maryland

(E) COUNTRY: USA

(F) ZIP: 20850

(v) COMPUTER READABLE FORM:

(A) MEDIUM TYPE: Diskette, 3.50 inch, 1.4Mb storage

(B) COMPUTER: HP Vectra 486/33

(C) OPERATING SYSTEM: MSDOS version 6.2

(D) SOFTWARE: ASCII Text

(vi) CURRENT APPLICATION DATA:

(A) APPLICATION NUMBER: Unassigned

(B) FILING DATE: June 23, 1998

(C) CLASSIFICATION:

(vii) PRIOR APPLICATION DATA:

(A) APPLICATION NUMBER: 60/050,667

(B) FILING DATE: June 24, 1997

(viii) ATTORNEY/AGENT INFORMATION:

(A) NAME: Brookes, A. Anders

(B) REGISTRATION NUMBER: 36,373

(C) REFERENCE/DOCKET NUMBER: PB387PCT

(vi) TELECOMMUNICATION INFORMATION:

(A) TELEPHONE: (301) 309-8504

(B) TELEFAX: (301) 309-8512

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 14063 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

AAGGTTGTTT GTTGATAATC TCTGCCATAT TACTGTTCCC TTCTTTTCGT TCATCGGGTA 60

AGAGCCGTCA GCGGTGAGCG CGCCACcTCC TTCACTAATC ACcACGTCGT GAACAGGCGC 120

TGCcGTAcAG CCGACCACCA ACGGCTCTTC TACCCACTCA ACTCTACATT CCACGCTAGC 180

GAGCCAACGC AGCAACGATC GTATCGATAT TGTGTGTTAC CATCCCTACG TAGGTACCCT 240

CGCTCGTACC CGCATCCCCC ATCGCATCAG AAAACAACTC GCCTCCAATn CTGnTACGTG 300

CCCTCTTGCC TGCACCGCAT CCCTTAACGC TTCAACGTTT TTGTGCGGAA TAGAACTCTC 360

AATAAAGATA GCAGGGAGTT mTACnTGCGC AATAAACGCT GCCAGTTCCT GCATATCATG 420

CGCACTGGCT TCCGAAGCGG TGCTCACCCC TTGCAACCCC TTCACCTCAA AACCATACGC 480

ACGGCTAAAA TAGCCGAACG CATCATGAGC GGTCACCAAC ACACGCcTTT CAGCAGGCAG 540

CGACTGCGCC TTGCGCCGAA CGTACGCGTC AAGCTTATCC AACTGCTGCT GGTACGCCTG 600

ATAACGTTGA GTAAATTCGC GAGTTTTTCC CGGCAACAGC TTGCACAAGC TTTCGTACAC 660

TGCCTTCACC GAATAAGACC ACAGCTTTAC ATCAAACCAC ACATGCGGAT CGAACTCTGC 720

TTCCTCAAGA GAAAGACGCT GAGACACCGG AATAGTCTCA GAAACTGCAA CTACCAAGCG 780

GCTCCCGCGC AGTTTGGAAA ACACCTCGCC CATCTTGGTT TCCAGGTGCA ACCCGTTGTA 840

CAGGATGAGA TCCGCATTCC CGAGCCATTC CACATCCCCC GCAGTAGCCG TGTACAGGTG 900

CGGGTCAACA CCAGGACCCA TCAACCCCTT TAGATGCACA TCACCTTGAG CGATGTTTTT 960

GACAGCATCC GCTATCATGC CAATGGTGGT GACAACCAGG GGTTTCCCGT CCGCTGCGGC 1020

ATCCTTGCTA CCGAATGCGT GCGTAAAACC GGTCAGCATG CCAAGCGCGA GCACGCAGGC 1080

ACATATTCTT TCACGTATCA AGTGACACTC CTTGGGTGAA TTTgATGCAT CAAAGTAGCG 1140

GATCACGCAG GAAACGTCAA TTTTTCAGTA CCcTTTCAGG AAAAGAAAAC GGCACTCTGG 1200

CGCGGACCTA CTCGAGCTAC ATGATAAAGA AGCATTTGAT CTTCCTCGCA TAAGCAGCTA 1260

CAGTTGCGCC CTCTTATGGA TCACACCGCG TCACTGAGTC CTGTGCGCCC TGAAGCACAA 1320

CCTACAGACG ATCGCGAGCG TGCGCTCAGA CCGCGCCTCC TGAAAGACTT TCTAGGTCAG 1380

GAGAAAACAA AACGCAACTT ACGTCTTTTC ATTCAGGCAG CGCGCGATCG CAACGAAAGC 1440

TTAGATCACC TGTTCCTCAT CGGCCCCCCG GGGCTcGGCA AAACGACGCT CGCGCATATC 1500

ACTGCATGCG AGCTGGGCGT TGAGTGCAAG GTTACAGGCG CACCGGCGCT TGATAAACCA 1560

AAAGATTTAG CGGGTATCCT CACTGCGCTG AGTGAGCGAA GGCnTTCTTC GTGGATGAAA 1620

TCCACCGCCT CAAACCAGCC ATAGAAGAGA TGCTGTACAT TGCCATGGAG GACTACGAAC 1680

TGGATTGGGT TATCGGTCAG GGACCGTCCG CGCGCACGGT GCGCATCCCA CTCCCCCCGT 1740

TTACCCTCAT TGGTGCAACC ACTCGCGCGG GTATGGTTTC AAGCCCGCTG ATTAGCCGCT 1800

TTGGAATCGT AGAGCGCTTC GAGTTCTATA CCCCTGAGGA GCTTGCTGCC ATTGTGCAAC 1860

GCTCAGCGCG GCTTCTAGAT ATCACGCTCG ACGCACGCGC AGTTnAGCCC TTGCGCGGTG 1920

TTCGCGAGGA ACACCCCGGG TGGCCAACCG GCTTTTGCGC CGTATACGCG ATTTTGCCCA 1980

AGTTGCGGGG TCTGCACACA TCAGCGAGAC GATAGTACGC GCAGGcTTGC CCACCTAAAG 2040

ATCGACGAAT TAGGGCTAGA ACTGCACGAC ATACAGCTGC TGCGCGTCAT GaTTGAGCAC 2100

TTCGGCGGAG GGCCAGTGGG CGCAGAAACG CTGGCGATCT CCCTCGGGGA ATCACCGGAA 2160

ACACTTGAGG ATTACTACGA GCCCTACCTT ATCCAAATTG GGCTCATGCA GCGCACCCCC 2220

CGCGGGCGCA TGGCCACCGC GCGTGCCTAT GCGCACCTAG GTCTCCCTGT CCCCGAGGCA 2280

CGCACGCTCA CCCCGCACTC CCCAGAACAA GGAACGCTTC TTTAGCAAAG ATGCGGACAC 2340

CTTGTCAGAG CTGTCGGCAG tCCGCTCAGA CCGGGTAAGA CACAAGAGTC TGAAAAAGGC 2400

ATATACTACG CGCGGGAGGG GTGCGTTTCG TGAAGATGTC TGCGTTTTTT GCACCAACCt 2460

GCGGTCTGCA CCTGCTGATG CAACCATCGC AAGCCACCAG CTGCTCATGC GCGCAGGGTA 2520

CGTCAGAAAA ATCGCCAACG GCCTGTTTGC GTACCTTCCC CTGGGCcTGC GCGTTCGACA 2580

CAAAATTGAA GCGATTATTC GGGAAGAACT CGAGGCTATC GGGTGTTTGG AGTGCACCGC 2640

GCCTGTCGTG ACTCCTGCAG AGTTGTGGAA GGAATCTGGC CGCTGGTACC GCATGGGCGC 2700

AGAGCTTTTG CGCGCCAAAA ATCGGCTCGA TCACGAGCTC CTTTTCAGTC CGACTGCAGA 2760

AGAATCCTTC ACCGCTTTGG TGCGCGGCGA CTGTACTTCC TACAAACATT TTCCCCTCAG 2820

TCTCTACCAA ATCAACGCAA AATATCGCGA TGAAATCCGT CCGCGTTACG GACTGATGCG 2880

CGCGCGCGAG TTCACCATGG CCGACGCCTA TTCTTTCCAC ACAGACTGCG CATGCCTTGC 2940

GCGCACGTAC GAAAAGTTTG CGCACGCGTA TCGCGCCATT TTCCGTCGCA TCGGCCTATC 3000

AGTCATTGCA GTACATGCAC ACcTCGGTGC GATGGGGGGG CAGGAATCCG AGGAATTCAT 3060

GGTAGAGTCC GCGGTGGGCG ACAACACGCT CCTGTTGTGT CCCCACTGcA CCTACGCTGg 3120

CAAATTGCGA AAAGGCCGTC GGACAGCGCC CCCTCCCAGA CACGCATGAC ACTCATCTAA 3180

AAGACGAACA CGAAGGgTCA GATCTCAAGA CGCCTGCAGC AATGCGCGAG GTGCACACCC 3240

CGCACGTGAA AACTATTGAG GAACTTGAAC ACTTCTTGCA CGTACCTGCA CATCGCTGCA 3300

TCAAGACGCT TATTTACCGC ATTGACACGG TGCCCCAGGC GGCTGGGCAT TTTGTGGCAG 3360

TGTGCATCCG CGGCGACCTA GAACTCAACG AGTCAAAGCT CGAAGCGCTC CTGCGCGTGC 3420

CATCTGTAGT ACTGGCAACT GAACAAGAGG TGTATGCACT CAGCGGCACC CCCGTAGGAT 3480

TCATTGGTCC GGTAGGAcTT GCACAGCGTG CTGCAGCTGC GTATGCCGCT CGCACCCtGC 3540

GTTCTTCCCC TCCGCTGCTG AGCCTGCATC CGTCACTTCT GACATTCCAT TTTTTTCCCT 3600

CGTTGCAGAT CAGTCCGTGA TGGCTATGCA CAACGCTATC ACCGGTGCGT TGAAAGTTGA 3660

CACGCATCTT GTGCAGGTAG AACCGGGTCG AGACTTTGTT CCTGACGCAg TTGCAGATCT 3720

CATGCTCGTG CGCGCCGGCG ACCGGTGCAT ACACTGTGGA GCGCCCCTAT ACGAAAAAAA 3780

GGGTAACGAA CTAGGTCACC TCTTTAAATT AGGGGACAAA TACACGCGCA gcATGcACCT 3840

TACCTTTACT GATGAGCAGG GTGTACGACA GTTCCCCCTG ATGGGCTGCT ATGGCATTGG 3900

CCTTGATCGC ACGCTTGCCT CTGTGGTGGA AAACCACCAT GACACGCGGG GTATCAGCTG 3960

GCCGCTTGCG ATCAGCCCCT ATGCAGTTGT GCTCATACCC ATCCCTCACA CGCAGGCCCC 4020

CTATGCAGCA GCAGAGGCAC TGTACGTGCA GCTGCGGACA CGGGGAGTTG AGGTACTGTT 4080

TGATGATCGT GCAGAGCGAC CCGGAGTAAA GTTCGCAGAC GCTGATTTAA TCGGTATTCC 4140

CTTCGTGTGG TACTGAGTGC GAAAAnCTAC CGCGCGTTGA ATGCaCAACA CGGTGTGGTG 4200

CGCACACGTA TTTTTTTACG CAAGAAGAGG CGTCCGAGCA CATTGCACGC CTGCTCGAAC 4260

AACTCGCTTC CCCGGAAAGT TCGTAAGAAC GGGAATGCCG GAGCGGGATC CAGCGCATGC 4320

AGTGCTGAGA CCTGCGCATA ATAGCACAGT GTACGGCACC CGTGGTTTAG AAAAAAATGA 4380

CGAAGGAGAA AAGGAAAAAC GGTGTACATA AAGGTAGCGC TCGTGTGTCT TTTCAGCATG 4440

GGAGCGCGGT GTCTTTTGGC CACAGAACCG GCGCCAGTCT CTGGAGATTA CGTATTGTAT 4500

CGCGACTATT CGTGGAAATC GCCCACATGG GTTGGCTTTT TGTGCTACGA CGCACACACG 4560

TACGGTGCGC TGCTGTGTAC TCCGGCAGAA AGCCGCAGGA TCACAATTCT CTTCACGGGT 4620

ACTGAAAAGC ACGGCCGCTT TGAGCTGACC GGACAACGCA TCACCTCACC GGTGCGCACA 4680

GAGGATCTGA CTGGCATAAA TTATCTCATG GATCTTTTTC CTCAACTACA GCGCTGGAAG 4740

CATTTTCCCC GGGATACACA CACCCTTGTT GCGCGGCATA CCGATCGGAG TAAAAAGAGC 4800

ACACAATTCT CAGGGGCAGT CGAACTGCAG TTCGCTTCTT TTGTCCCCCT CTTCCACCTA 4860

GAAATACTCC GTGATAAGCA GCAGCGCGTC ATGCTCCAGC TAAGCGAGAT AGGGAAGATC 4920

GACCACACCA GTGACGCAGC CTTCTTTCAA TTCACCCCCA TGCCCCCGTC CACGCCCACT 4980

GATGCACCGc CAGCAACGCT TAATCAGACC CTGACACGCA CGGAGTATGT CATCGATGAC 5040

GTGTGCATTG CACTTGATCC GCAGTGGAAA AGAATTGCAG AAAATTCCTT TCTTTCAGAC 5100

TTTGCCTTTC TCACCGTACA CCAGGTGCCT GCACCGCGCG CGCACGACTA TTCTGCGCTC 5160

CGTGCATTGC TGCAACTCTT TCTGTATTCA GGCCCTCAGG GAAAAAACAT TCTTGAACAA 5220

CTCCATATCA ATGACACTCA CGCGCGTCTT ACGCTTTCCT ATGCAGTGTT TGACCTTCCG 5280

TCAAAAACAG TTAAAAAGAC ATGGAAGATA TTCATCCGCC ACTCTGATAC GCACTACTCT 5340

ATACTTAGTC TCACGGCGGA CCAgCGCACA GCGCAGsGTT ACGCGCGCTA CTTTGACACG 5400

CTCATTGAAA CTATCCGTAC AAAAAACTAA AAAATGCTGA ATTGGAGCAT ACCCGTGATT 5460

AGACACATAT TATTTGACAT AGACAACACG CTGTACTCCT GTACAAATCC CATTGAAATG 5520

GCTATCACGC AGCGCATACA CACATTTGTT GCACATTTTC TCCACGTATC TTGTGAGGAG 5580

GCGCGCGCGT TACGCCAGCG CACAAAGCAC CTCTATGCTA CCACCTTTGA GTGGTTAAAG 5640

GCAGAGCACA ATCTCATTCA CGATGAACAC TACTTTCGTG CCGTATATCC TCCCACCGAA 5700

ATACAGGAGT TGCAGTACGA TCCGATGcTC CGCCCTTTTT TACAGTCACT GCACATGCCA 5760

CTGACGGCAT TAACTAACGC ACCGCGCGTG CACGCACAAC GCGTATTGGA TTTTTTTCAT 5820

CTGTCAGACC TTTTTTTAGA TGTCTTTGAC ATCACGTATC ATGCAGGCAA GGGAAAACCA 5880

CACCACAGCT GCTTTGTACG TACGCTTGAA GCGGTACACA AAACTGTGCA GGAAACGCTT 5940

TTTGTCGATG ACTGTCTCAT GCACGTGCGT GCcTTTATTG CGCTTGGCGG ACATGCCGTG 6000

CTGGTTGACG AACGTGACTG TCATGCAGAA CTGCCTCCTT CTGCACGCAT GACACGCGTA 6060

AAAACAATTT ATGAATTGCC CGCACACCTT GCACGCCTCG CCCAAGGAGA CAATCAGTGA 6120

GTATACATTC GTTGCAGCAG ACTTTTAGCG ACATCGTCCC GCTCCTGGAG CAGTATACGC 6180

GCGCAGACCG CTTCATGCGG GAGGATAATT TGTTACACGA GAGAAACGAA CCTATCCGGC 6240

GTATCGTTGA GTCCCTCGTC GCCCGCATAT TACTCCCCGG CTCCACAATG CGCGGAAATG 6300

AGCAAATCGC ATCCTTTTTA CATAAAACCA ATGAAGGGAA ACGGGGACTC ATTCTTGCGG 6360

AACACTACAG CAATTTTGAC TTACCCTGTC TGCTCTACCT TATGGAACAA GGAAGTAGTG 6420

CCGGGCGCAT GCTTTCAGAA AAAATCGTAT CTATTGCCGG TATTAAACTT CGTGAAGAAA 6480

ATCGCATCCT GGCAATGCTC ACCGAAGgAT ATGATCACCT GGTGATATAT CCCAGTAGGA 6540

GTTTGGCCAC CATCACTGAT GCGCACTGTC TTGCAAGAGA GACAAAGCGC AGCnGAGCAC 6600

TGAATCGTGC AGCTATGAAG TATTTAGAGG AACTGCGCAA CGCGGGAAAG GTGATTCTCG 6660

TGTTTCCTGC AGGGACACGC TACCGACCCG GGAGACCGGA AACAAAGCGA GGGGTGCGCG 6720

AAGTATACTC CTACATAAAA CACGCCGAGG TACTGCTCCT TATTTCAATC AATGGGAATT 6780

GTTTGCGCGT TGCAGAACGT TCAACTGATA TGACGGAAGA CGCGGTGCAT CCGGATGTCG 6840

TGCTTCTTGA AGCGCGCACT GTAGACGAcT GCGCCCTTTT TCGAGAAAAA GCGCTGGACT 6900

GGCACCGCAC ACACAACGTG GCGGCACCGT CAGAGGATAA AAAACAAATC GTAGTCGACT 6960

ATGTCATGCA CCTTTTGGAA GAAATGCACG AGCACAATGA ACGAGAAAGG CTATCGTGAA 7020

TTTTTCGCTG GAATTCCCCG TAAGATCCTA TGAGCTAGAC GGATACGGAC ACGTGAACAA 7080

TGCGGTATAT CTCCAATATT TTGAATATGC GCGCGCCGCT TTTTTGCTCC ACATAGGGTT 7140

CGACCTCAAA CAGTTGCACG AAGCAGGTTA CGCTTTCTAC GTAACCCAGG CGCACATTCA 7200

CTAcCGCACT GCAGTGCATC TATTCGATAC GTTGCGCGCC CGGGTAAAAC CATTAAAGCT 7260

CGGAAAAGCT TCCGGCGTCT TTTCACAGAC GCTGGAGAAC CAGCATCACG TGCTATGCGC 7320

GGATGCGGAA ATTACCTGGG TGTGCGTTTC GCGCACAAGC GGCAAACCAA CTAAGATTCC 7380

CCCCGAGTAT CTGGTACCTG CGCTGTATCC GAACTACTAG TCCTCCCTTC TTTCCCCTTT 7440

ACTCTCCCAA GGACATCACA CTACGGAAGG GTACGCATAC GCAGTAGGGA GGTAGGGTTT 7500

ATCGCGGAGC CATTCTTATA GATTGTAAAA TGCAGGTGTG GTCCCGTGCT GCGTCCTGTT 7560

TTTCCCAATA ATCCGATTTT tGTCGCGCTG GTGACGCGCG TACCTGCTGA AACCAACACC 7620

GTCTGCAGAT GCCCATACAG GGTCTGATAC CCCGCGTGGT GCCCCACAAT CAGGTAATTA 7680

CCATACACTG CACTGTATCC AACCGTGCGT ACAATCCCTC CGAGCGCCGA ATATACTGGG 7740

GTACCCCGCC GACTCACCAT ATCCAAACCA TTGTGAAAAC TTCTGGCACC GGTAAACGGA 7800

TCAcTACGCC ATCCATACCG CGAAGAAACA TAGTACCGAC TGCGAAGAGG AGCACGAAAC 7860

AAGTCACCAT TAATTTCCTG CAACGCGCGT GCGCTTAAAT GTGCACCGGG CAAAAACAGT 7920

ACGCGTGCAG GCTGCAATGG CTGCAcTGCG TCAAACGACG TATTTTCCCT CCACTGTTTC 7980

GCAGAAGAAA ACGGAAAAGG CACGCAGgAC TCCCGTGCAG CTGAATTATA GAACGGAGAA 8040

ACCAGCGTAC GCACTGAAGG AGGTGACTCC TTTGAAGAAG ACGGCGTGTT AAGCAGCACC 8100

AATCGTTCTA AGGAGATCTG ATGCGCCGCC GCTATAGACG AAAACGTATC GCCGTTTTTT 8160

ACGGTATATA AAATGCCGTC CACTGAGGGG ATTTTTAGTA GCTGTCCAAC TTGGAGCGCC 8220

CGTtGyTGCG CAATTTATTC AAACTAATGA TTGCATCCTG ACTGATGTCA TAgcgCtGCG 8280

CAATCCTTCC TACCACATCA CCTTCACGCA tTCGTACACT GTGTAGTACA GTGCAGGcTC 8340

CGCATCTTCC TGCACGATAC GTGCACGGAG CAAGGAAGAC ACGTACCCCG ACGCCTGACG 8400

TGGTTCCTGC TCAGTGAGCG TGAGGGCAGG TGTCAATGGT TCCACCTGAG CACCAAAGTA 8460

CGCAAGGGCA AGAGCAAGGA GCAACAGTGT TACGAACAGT AACAGTnGCC tACGGGGACA 8520

GGTCTACACG GTTCTCGCAC AGTCTGTTTG GAACTTCGAC AGTACACGCT CACACCGGCT 8580

ATCCTTCAGG TGTACACACT GCCGTATCtG CGGGCAGGTT GCGTCTGTAC CTAACGCACC 8640

GTCTAGAGCG TCCACGCACG CAcGCGCGCG CGCGAGGGAG TCCGGCGGAA AAGAAGTTAA 8700

CACCCGCATG AATGCAGGGC TCGGTGTTGT CCACACCTGT GCAGGCAGCG CCTGAAGACG 8760

CGATGAAGGA AAACGCACAC ACCAAGCAAG CAAAAAAAAG nGCGTGTAAC GCGCATTCCC 8820

GTGcTGACTC GCGCCACCGT ACGTGCGATC TCCTACCAGG GGGAATCCCT GTGCAGCGCA 8880

ATAACGGCGA ATCTGATGCT TTTTCCCcGT AACCGGCACA ATCACGCGGA GCACCAGCGC 8940

GCTATCACAG CTATGTAACA CTGTTTGCAC ATGCGTTACC TCTCCTGGAC GCACCAGCGT 9000

GCGCGCCGCA GCAGCGGTGC gCGCAGGGGC GGCGGTGATC GCAAGATAAA ACTTGCGCAA 9060

TGTATGCTGC TGCAACGCGG CAGAAAACCA CTGGGCACCG CGTAACGAGC GCGAAAAAGC 9120

AATCAGTCCC TCTGTCCCTC GGTCCAAGCG GTGCAACGGT CCAGGGCGGA ATGACAAAGC 9180

AGGGGGAACG TGCGCACGCC CTTGTCCCCT CACCCAGGCA TCCAGGctGC GCGGACCGTG 9240

CACAcAcAAC tGCGGGTTTA TGAAAAAAAA GCAAATCTTG TGTTTTAAAT ACCACCGAAG 9300

cAACACGCGC ATTcGGTGTT CCAGGCATCT TCGAAAGACG ACTGGATGCT GCACGCGCCC 9360

TACACAGGGA TTCAGGTAAA GAAAGCACAT CCCCCACCTG CACCCGCTTT GCAGGCTGCA 9420

CCGGACGACC ATTGAGCCGG ATAGCGGTGC GGCGCACGCG GGCATACACC CCAACACGCG 9480

GACAGGCAGG CAACAATATT CGCAAAACAC GATCTACTCG TCTACCTGCA TCGTTTTTGG 9540

TGCAGCGAAA ACACTCAACA GCCGCTCCAC CATGGGGTCT CACGGTGGAA ACAGGAGGGA 9600

CGACATCCAT ATGCACAGTG TGGGGAACGT TAGACGAGAC CCACCTTTTC ACGCGACGAA 9660

CACCTACTTT CATACACGGT GCGTCCCCGT GCCGGATACC AGTTGCTCCT CCCAAACGTC 9720

CTCCCCGTCT TTCCCAGTAC GACACCACAG CCCATGGCGG GTACAGCCGC CCCAGTATAG 9780

CGCACACAAC GCTCCCTTGA CAAAGGTTTA GAGAGTATAG GAGACTGCCC CGCGATGGAC 9840

GGTGGCTATT TTCTTGGCCA GCTGCATGCG GTGTTCAGTG GTGAAGTCTT CCTCTCTGCC 9900

ACCTGTAGTT GGCTTGCAAG TCAGGTGATT AAAGTGGCTA TCGCATGCCG AAGtCGGCTA 9960

TACGGTCGGT GCACGGCTTT TTTGATTTTG CTGTTTGGCG CACCGGCGGC ATGCCTTCGA 10020

GTCACTCTGC TCTTGTGTCG GCGCTCACGC TCTCTTTTGC GCTCAAGTGC GGGTTGCATT 10080

CGGATCTGTT CATCTTTTCC TTTTTCTCTG CCATCATTGT CGTGCGCGAC GCGCTCGGTG 10140

TGCGCCGTTC AAGCGGCCTG CAGGCCGAGG CGCTCAATAG CCTCGGTGCG CGTGTTTCGG 10200

AGAAACTTGA TTTTTCTTTC AGACCAGTGC GAGAGATTCA TGGACATAAA CCGCTGGAAG 10260

TTGTCGTTGG CGTGGCAGTG GGCATCGTCA CGAGCGCTTT GTTCTACAGC TCCATGAGCC 10320

CTTGAGTCTC CGGTGGACGT GCATGCAATG CGGcGGACCC CTCCACACAG AGGAAGAGGC 10380

GGTGCTGTGC GCGTTCTCTC GTGTGTCCCT GCCGCGTCGG GAGGCGCAGA CCTTTTCTGT 10440

ACCGTACAGA GGGCACACCA ATGATAGAGC GCCTACGGAG CAGTCGCGGG AAACTCACCC 10500

TCACCCACCA GATTTTCCCC CTCAGCTTTG GGGGGAATGC TTTTTTGCCT GCGCGCGCGC 10560

TCGTTCCGTT CTCCGTTGAT GCTGGAGAGC CAGCCGCCGT CGCTGTGGTA AAGGTTGGGG 10620

ATACGGTCCG AGAAGGTCAG CTGATCGCAC GCGCCGCGCA CGCCGGTGCT GcTCACGCAC 10680

ATGCCTCCGT CCCCGGTGTC GTCACCCGCT TGGTAAGTGC TAATTTTCTC GCCGGTAGTG 10740

CCCTGCGCGC TGTCGAGATT CGTACACGCG GTTCCTTCGA ACATCTTGGC AAGGTCCAAC 10800

CAAATCGCCC GTGGCAGCAC AGCACCGCTT CAGAATTGct GCGCCTAGTT ACAGATGCAG 10860

GAGTAGTGGC CACACGCCTA CATCCGCACG CCCAGATCAC GAGCACCGCA ACGGGCACGC 10920

ACGCGGGTGC ACAGCACACG TACGCGAAAG ACTACGGACA GAAGAGAAGG GCTGAAGCGC 10980

ACACGCTGCG TCTCATGCGC GCGGCGTGGG AAAGCGGCAA TGCGCTCGCC ACGCACCTCC 11040

ACCTGCACGT GCGTAAGGGT GTACGGAAAC TTACGCTCTA CCTTTGTGAC GACGACGCTA 11100

CCTGCCCTTT GAGTTCGTTC CTTGCGCAGG AGTTTCCAGA ACCTGTTGCT ACCGGTACCG 11160

CCATTATTGC ACGGATACTG GACGCTACGT ATACCcGCGT GTCTCCACAC GCTGCCAAAA 11220

CGCTCCCCCG GTCTTGCAAG GATGCGCGCT GTCTTTCCAT TCAACGAGAT GCACGACGCG 11280

TATAGACGAC ATTATCCTTT TAGCAATCTA TGTGCCCCAC GCTATCGTGC AGGTTGCACA 11340

ATCGATGCAC TCACTGCAGT GCACGTGTAT GAGGCAGTGG TACTCAGTCA GCCGCAAATC 11400

AGTTCCTACA TTGCTCTGAC AGGCGCTGGA TTAAAATCAC CGCAGGTACT CCGCGCGCGT 11460

ATCGGCACCC CCCTTGGCGC GCTCATCGAG GAGTGTGGAG GGTTTCGCAC ACGCCCCGGG 11520

CATCTCATCA TCAATGGACT GCTCAAGGGT AGTGTTTTAG AGTCGTTGGA CCTGCCTTTC 11580

TCAAAGGGGA TCAAATCGCT CCACGTCACC GGTAAAGCGC TTTCAAGCTC TGCGTCCTGT 11640

ACCTCCTGTC AAAACTGTGG TGATTGCGCG CGCATTTGCC CAGTATATCT TGACCCAATA 11700

AAAATTGCGC GTGCCGCACA CCGTAATCAG TTTACTGAAG AAGTGCTCCA ATCCcTGcGG 11760

ATTTGCCACC AATGCGGTCT GTGTTCTGCC GCCTGTACTG CGCGTATTCC TCTTGCAAAA 11820

CTTTTGCACG ATGCACAAGA ACGCGCACTG CATCTTTCCC GTGCTCCAGT CACCAAAATA 11880

GAACCCCACT CCACACAAAG CGTCGGGAAA ACTATCCGCG AGGCACCTGC CAATGCGCAC 11940

CGCTGAGTAC AAACACGCAC CCTTCCTTTA CACCGGCTTA AGTGCTGGAC AGAACAACAG 12000

TGTACTGTTG GCGCTGCTTG TTGCGCACGT GTTCGTCGTT GCAGCcaTkc gCGACACGGT 12060

CGcGCTTTTT TCCATCGTCA GTACCGAACT CGGCGCACTG AGCGCCGCGC TCGTTCAAAC 12120

AcTACGCACA CCACATGTGC CCCTGAGCGA CTCTCTCGTA CTGGGCCTGC TCATCGGTGC 12180

AGTACTCCCC GCACACAAcT CTTTTTTGAA cACATTTTGT GTCGCGTTCT GTGCCgTATT 12240

TTTTACGCGC GTTTTGTTCG GTGGCAAAAT CGGGAATTGG CTCAACCCCA TAGCGCTTGC 12300

CCCTGTCCTC CTCCGTCTGT GCACGGAGGG AACTTCCCTC CCAACGTCTG GGCGTGTCTC 12360

TGTTGTACAG GGAGCGATGT CTTATCCTCT TTTCTATTCT GCGCTTGTCG AGTGGGACGC 12420

CGCCGTGCGT ACGTGGTGCA ATACGCAGGT GTTCCAACCA CTTGGTCTTA CCCTCCCTGA 12480

GGGAGCGTTG AGCGCCTGTG TGTTCACTCA GGCTGCAGCG CCTGGGTTTC GCTATCCAGT 12540

ACTTACCCTT CTTGCTGCAC TGTGTGTATA CGCAgTGCGG GCGCGACGCT ACATCTGTTC 12600

GTGCGCGTTC CTTGTGGTGT ACAGCACACT GTTTTTTTTa CCCGCACACG CACACCCTGC 12660

AmCCCTTGTT TCCCTCATAA AAAGCGGCGC GCTGTTTACT GCATTCTTTG TACTCCCTGA 12720

GCCAGATACG TCAATGCGCA CAAATGGCGG GGCTTGGATC TCAGGGGGAC TCTGTGcTAT 12780

GTGCGCGTTT TTTCTTGCAA AAAAGAATAG TTCTGCCCCA GATATGTGGG GTGCACACGA 12840

CATGCACTTG TGGAGTGCGA TACTACTCAC CAACATCGTA CAGCCACTCA TTCTACGCGC 12900

AGAATCCTGG TACTACTATG TGCGGAGGCG TCGCTATGAC GTACAACACT AACACGAGTC 12960

TTTCATCCTA CGCAGgATTG AGCGCATTTG CGTTGTCAGT CTTTTGCATT CTATGGGGCA 13020

CCGCGCGCAC TGGTTCTTTT TTAAAAGAAA AGGCGCTCAT CACyTGCGCC GCAGATATCC 13080

TTGCAAGGCA AGCCCCAGAA CTTGGGGTCA CGTCACGCAC CCTGCGCATG GTACCGAGCT 13140

CCCCCATACC GCAGgCTGAG GTGCTTCGGG GAAAAAAGAA TACGGGAGAG GAAATATTCC 13200

TATACTTTTT CCCACTCAGG GGAATGTACG GTTCGTTTCC TACCCTTTTT TTGTACGATA 13260

AAAAAGATGG TGCaCGCTTT TGCCaTCTCA TAGGTAATCA CCCTACACCG CGTGATGCAC 13320

GCTTTTATGG CATATCGAgT tGCGCGCATC GCkyTTCAGT GTAGAAAAAT AGAACACCTC 13380

CATCAAACAG TCGCATATGA GTAAGTACAC GGTTAAGCGC GCGAGTGTAT TGTGCATTTT 13440

TGGCATAGGA CTATTTGTTC CTGCAACCGG AACCTTTGCc TGCGGTCTAC TACTCGTACT 13500

TGGCTTTTGG GTTCTATTTT TTTCCTCGCT GCTGGCGAGA TTTCTCTCAC AGTTTTTTAT 13560

GCGCACGCGC AGCgcTCCTT TGTTCGAGGT CTGTCTTACC CTCTCAGCCA CCATTATGTA 13620

TGACAACTTG ATCCAAGGCT TTTTCCCGCT TGTGCGTATG ATGCTGTGTC CTTACCTTTT 13680

CATTAmCsCG CTTTCGCGCA CACTCGATCT CTGTCTTACC GCATACGATG CAGATGCCGA 13740

ATCGCTCGAA TGCGTAGGTG TCTTCGGCAT CATGATTGCG GGAATTTCTC TTGTACGTGA 13800

ATTAGTTGCC TTCGGGTGCG TTTCGCTACC GGCCCCGTCG GGGTTCTTGC GCATCATCTC 13860

TTTTCCACCC AGCAATGTAA TACGCTTTGC AGCCACCGGC GCAGGGACCC TCATAAGCTG 13920

TGGTATTGTT CTTTGGATAT TCCGCAGTGC AGGTAACGAC CACACGCCCT CTTTAAGGAG 13980

TGAATGGTGA CAATGGTGCC ACCCCTGTTT TTCGTATGCG CCCTTTTCTT TGCCGAGGGC 14040

ATCGGATTAG ATCGCCTGGT AnC 14063

(2) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 14244 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

CTCTGCTTGC CCCCTATGAG AAGACGGAGG CGCTTTCTCA CTCTTTGCGC GTGTGCTGTG 60

CACCTTCTTC CTCTTTTCCC TCAGACGATT ACAACCGCTT CTGTCTTTTT CGCCTCAGTT 120

TCTGAGCATT ATGCCCGGTT AAAGGATTAC GCTGCTGATT TGGCCATGAG CACCGGGTCA 180

GGAACCCGCG CGCACCTTAT GCGCGCAAAG GTTATTTTTA AATATCCAGA CCGTCTGCGT 240

TTGGATTTCT CAAGCCCTGC TGAACAAACT ATTGTCTTCA CGGGAGATAG CCTGACCATC 300

TACTTGCCCA CCTCCCGCGT CGCGCTTGTA CAATCGGTAG CAAAAGATGA CACAGTAAGT 360

GCTGCTTCTC TAGCTTCGCC TCATGGTCTT GCGCTTATGA AGCGGTTCTA CACGATAGCC 420

TACGAGACGA GTTCTTCTCC TGTTCCCCTG GGTCCGGACA GTGGGGAGAT GGtCGTTGCA 480

CTGGTGCTCA ATCGTAAGTC TGCAGCAGAA ACATTTAAAT CTTTGCGCGT GCTTGTCTCG 540

GCACATACCA AGCTTATCCG TCGCATTGAA GCGTGGCCTC TTTCGGGGGA AAAAATAACA 600

TTTGATTTCA GCCACTATCG TTTGAACGTC GGTATTCCAG ACACACGGTT CCTCTACGAT 660

GTGCCCCCAA CCGCAAATGT GGTGCACAAT TTTCTCTTTG CTGATTGACC GCTGCCCCCA 720

AAGGACGTGA CGATGCCAGA TATTGGAGAG CTGCTAAAGA CGACGCGCGA ACGCAAACAC 780

CTCAGTCTCG AACAGGtGCG CACGAGACGA GTATCGCACG CCGTTACCTG GAGGCGCTCG 840

AGAACGATGA GTATGATGTT TTTCCCGGCG AACCCTACAT CCTTGGCTTT TTGCGCAATT 900

ACTGCGAGTA CCTCCAGCTG GATACGGAGC AGTGCATCGC TCGCTATAAA CATTTAAAAA 960

TTCAAGAAAT GTCGCTGCCA ACGGAGACCC TCCTACCGAG TAAACGGTGG GGTTCATTTC 1020

CCCTGtTAAA rGGAGTTGCC TGTGTGCTCT TCCTGGGTGG GGTGCTGGGT GTGTATTACG 1080

CGCGGCACCG CnCnTnGGGT TTTCTATCCC GtATTGTGTT CTTTGGCAGA GCACAGCGTA 1140

CCCCAAGGGA GCTGTCTCCC CCCGATGCAA CGGGGGCGGT GCGCGAAACA GTGTCGCTGT 1200

CTTCTGCACA ACATGAGGAG CGTGCGCGAC GCACCGTATA CGAGCGCATC TCGCTATACG 1260

CTTGCTGAGG AAAAGTTTGA ACACACGGTC TTTCCAGGAG ATGTGTTGGT TATCAGTTCC 1320

GGGGGGAATG CGTACGAGCT AACCGTCAGC CGCACTACGC CGCACCTGTA TCTGGACACG 1380

CCCATTGGTA CACAGGTGAT CTCTCTTGGT CAGCGCCtAG TGATGGATTT GAATACAGAT 1440

GTGCAGCCGG ACGTAGAAAT AAGTGTGGAA GACATTGAAG CACATCAGGC GGACGGGGGC 1500

GCGCkTGTTC GCGTGTTTAC AGGTaGTCTG GTGCAGACGC TCCGTGAtCG CAgTGCTCAG 1560

AGCTTTGTGC CTACAAGTGG GGTAAATGTC TCTGGTCAGA CGGGAGTCGC TGCCGGCGCG 1620

CGATATCAAG TTTTGTTTGA AGGCGGTGTT GCGTACCCGG TGACAATGAA CGCAACGTTT 1680

CGCTCGTACT GTTTGTTCCG GTACGAAGCA GATCGCACGC GGCGGGAGGA GCGGTATTAC 1740

CAAAAGGGCG AGCAGCTGAC GGTGCAAGCA AACAACGGGA TTCGGGTGTG GGCATCTAAC 1800

GGGAATGTGG TGCAGCTGCA AATTGTCGCA GGCGGTAAGA CGGTGGATGT AGGCCTCAGC 1860

CGTCCGGGGG AAGTGCTGGT CAAAGACATC AAATGGATCA AAGATGAGGA CGCCGGGCGG 1920

TTCAAGTTCG TGGTCATGGA AGTAGACTAG CGCGCGGCGG CAGCAATCGC GTACGCkTTC 1980

CAGAGCGCGT GGACTGCAGT GCACAGTGCG CTTGCgCGCG CGCGGGAGCC GCTTCTTTTT 2040

TTCTCTCTTA CAAAAAGTAC CCGTAgCGCT GCGCCCGCAG CTcCTGCAAA CAGCGTGGcG 2100

CTGCcTGCGG GCCGGTGTGC AAGAGCAAAG AGAAGGACTG ACAGTACCTC gCCACAGGcG 2160

CGTGCAGTGC AGGAAGTGGC ATGGTGGCAC AGACGCTCAG GTATATAGGC GCGAAAGAGC 2220

ACTTCTTCGC TCAGAGCATT TAAAAAAAGC CGTACATAAA AAGCTCCCCC TTCCCCTTCT 2280

GGGAAGGGAA ACGGTGAGGG AAGAGAAAAA CAGAAGGAAG GAAATACTAG CGTGCTGAGT 2340

ACGAACGCGT ATTCGGTAAC AGCCGCCGCA GCGTGnCAGT ACGTGTGGCG TGTTGCGCAA 2400

AGGGGAGAGG TGCATCAGCG TGGAACAGTG TAACAGAATC GGGCAGAGGG GGTACAGAGC 2460

GCATATATCC CCTGCAGTGG TGATGGCCAT TGCTGCGGTG GGAGGTTTTT ATGGGACTCA 2520

CGTGTATGAG GTACCGTTTC CGTATGCATT CTTTGGTGCA GTACAGGCGT GTGTGCTGTG 2580

TATTGGGTGT TTGTTGGTCC GCAGTGGTGT GCGGTTCTTT TCTCGTTGGG GTGCTGTCCG 2640

TATCTGGAGG AGGTGGGGAA TCGCATACAC CAGCGTATGT CGGTGTTGTA ATACGCTTTT 2700

TTTCGTGTTC TGTGGTCTGT GTGTTGCCTG CGTTGCGCGA ACCTCCCTCA TGGTACAACA 2760

AGCTCCGTTG CAAACACTTG CACAACCCCA AAAACTACGC GTTTTGACTA TACACCTTTT 2820

GCaAGAGCCA AAGCCTGCAG GCaCGCGCTT TCGTGTTCGG GCGCGCGTAT TGGGTGCAGG 2880

TTACATAGAC GGTGCTTCCT TTTCTGCACG TGGGGTGTGC ACTGTATTAT TTCCTGCAGA 2940

GGTAATTTTG CAGCAGTACG CTACCGATAT GACGGACGAC gcGGATGCCC GCGTCTGTCA 3000

GTATTACGCG CGTGGGTTGC GCTGTCAGAT TCGTGGGCGC TTTGCATCTT CTGCACCGAA 3060

GCTTTTTATC AGTAGTTCTA CACCACCACG CTTTGTTGGC TGGAGTTCCT ATTTTGCACA 3120

GATGCGCGCA CAGATGCGGG TTGCACTCAT GAGGTTTTTA TCTCCATGGG GGCGTGCAGG 3180

GGGATTGTTA CTCGCGCTCC TTTCTGCAGA TAGTGTTTTT CTTTCGGATG AAATGCGTGT 3240

CGCGTTTCGC CATGCAGGAC TTGCTCACGT GTTGGCACTC TCTGGCATGC ACTTGTCTTT 3300

GGTAGGGGCG AGTGCAACGT tTTTGGGCCG TTTCATCGGC ACAAGGCACA GAGGTATGCA 3360

GGGGGCGTTT TTTGCGATGC TTGTCTTTGT GTGGTTTGCA GGTATATCGC CTTCCCTTGC 3420

GCGTGCACTT GGTATGACTT TAGTGCTGAT GGGAGGACAG ATGGCATACG TGCGCGTAGG 3480

ACTTTTTTCT GTACTGTGTG CTGTACTTAG CATACATATG CTCATTGCGC CGCATGATGT 3540

ACAGACGTTA AGTTTCATGT TGTCATACGG AGCGCTTGCA GGTATTGTGT TGCTTGGCTC 3600

TGAGATTACT GAAATGATGT CGGGTTTGAT TCCTCGGCCA CTTGCATCGC TGCTTGGAAC 3660

GTCCTGTAGT GCGCAGTTTT TTACAGCACC GATAGTGCTT TCGGTCATTG GATATTTTGC 3720

CCCCATTGGG GTACTTGCCT CGTGTGTGGT TAGTCCGCTT ATCGCCTTAT TTTTGATAGG 3780

GGGGAGCGTG GCGCTGTGCT GCTCTTTGGC AGTGCCTGCT GTTGCGCCTT TTTTAAGTTG 3840

GGGTGTGTAC TTTTTTGGTG AAGGACTCTG TGCGGTTGTG CGTTTTTTTG CGTGTGCGCC 3900

GCTTGTGTAT GTACAGAGTG CCTGCGGACA TGTGTGTGCT GCATTATTTT CTTTTTTACT 3960

CGGTGGGGGA CTACTAGAGG CGGCGCGTCG CGTGcGTGTT CACAAGGATA CATATGTGTT 4020

GCCCGAATTA TAATTCGGCG CGTGCACTTG CACAGTTTTT GACGGAGCGC GGTTTGCGGA 4080

TGCATAAAAA GTGGGGGCAG AATTTTCTGC TCGATCCGGT GTTACGTACG CAGCTTGTTA 4140

AGATATTGGC GCCGGAGCGT GGGGAACGTG TATGGGAAAT tGGTGCAGGC ATTGGTGCGA 4200

TGACCGCACT TTTGGTGCAA AACAGTGATT TTTTAACAGT GTTTGAAATT GATCGCGGCT 4260

TTGTGCAGAC ATTGCGCAAA CTTTTTGATG CACACGTCCG TGTGATAGAA GGGGATGTGT 4320

TGCAACAGTG GCATGCTGCA GCAGCACAGG AACAACCTGC GTGTGTTCTA GGAAATTTAC 4380

CCTACAATAT TGCTGCCCGT TTTATTGGAA ACACGATCGA ATCAGGCTAT ATTTTTAAGC 4440

GTATGGTGGT GACCGTTCAA AAAGAAATCG GGTTGAGAAT GACTGCGCTC CCTGCACAAA 4500

AATGGTATTC ATACTTTTCA GTACTCTGTC AGTGGCAGTA TGAAGTGCGT GTGATTCGTA 4560

ACGTTGCGCC TGTCTGTTTT TGGCCGCGTC CTCATGTAGT TTCTCAAGCA TTGGTACTCA 4620

CCAAGCGTAA TGCGGTGCCT TCTTGTGTGG ATCCTGCGCT TTTTCTGCAC GTGACGAAAA 4680

CTTTGTTTTC TGCGCGGCGT AAAACGGTAA GAAATAATTy ACTCACGTGG CAAAAAAGGA 4740

TGCCAGGCGG TGCAGCTGTG TGTGTAGAAG AACTCTGCGC ACGTGCAGGT ATTGACGCGC 4800

GTGCGCkTGC AGAGCAACTG AGCATCTATG ATTTTATTAC GCTTTCTGaT aCgctGCGCG 4860

CGCTACTGTA GTCCGGTGTG GGTGTTGAAT GGCGCGTGTC TATATTCTTT TTTTCAGTGT 4920

GTTTTTTGTT TTTCCGCTCT TTTCTGAAGA CGCCGCGCGC GATGTGGAAC CTAGCGATGC 4980

GCCTGTGCCC TATGAGGACA CAGAATTTTC CTTATGGCAG AAAGAATTGT ATCGTTTTGA 5040

AGCGCTGTCC ATCGGTGCAT TCCCGATAGT AACGCTGCTC TCTTTTATCA CGTATGACAT 5100

CATACGTCTT ATTCAGCAAT GGTCGACAAA GCCTCCGACA TGGTGGGCGC TGATTATTCC 5160

TGGCGCGGAc TAcCGCCAcT GAGTACGAAG GAGCGCGCGA TAGTTTTTGG TGTGGCAGTG 5220

GGGATTTCTG TGACGATTGG ATTAATTGAC GTGACGTATC GTGCAGTGAA GCGTGCAATA 5280

CACCGGCGTA GTCTTGcaGC GTTCGCAGTT AGTACCAGAC CCGATAGAAC TGGTGCCACT 5340

TGATTCTTTT GTTGAGGGGA CTGACGATAG CACGTGAAGG TGCACAGGGT GTCGTATTCT 5400

TGCAGTGCAG AAAGCACGTC GGTGTGCAGT GTTCATTTTT CCGTTTTTAG AATACCGGGC 5460

GCGACGTGTC GGTGTGATGT GTTCTGCGCA GAACATGTTT TTTTTTGTGC ACGATCTCAG 5520

CTCAAGAGGA TCGTGCGTGC ATTGTGTGTG AATGGGCATA CGGCAAAGTT TTCAAGACCT 5580

CTTCACGTGC GCGATCGGGT GTCTTTTGAG TGGGTACGCT CAGTGCCCCC GGCGCTCATT 5640

CCTGAGAATA TATCGCTTTC TATTCTGTTT GAAAACGAAG ACATTATTGC GGTGAACAAA 5700

GCGCAGGGCA TGATAGTACA TCCTGGGGCA GGCCACTGGA CGGGAACACT TGTTCAGGCG 5760

CTCAGTTTCT ACCGGGTGTA TCGTGCACGT TTTGAGGATG AGTTTTCTCG TCAATTTCAG 5820

AAAGGATTTC CCGATTTTTT CAGTACCCTG cGTCAGGGTA TTGTGCACCG TTTGGATAAA 5880

GATACATCGG GCGTACTCCT CACTTCGCGC AACATGCATG CTCATGAGGC ACTTGTACGT 5940

TCGTTTAAAA AAAGACAAGT AAGAAAAGTA TATCTTGCGT TATTGCAGGG TGTTCCTGCA 6000

CGCGGGGTTG GGGTGATTGA AACAACAATC GTGCGAGATA GAAGACGACG CACGCGGTTT 6060

GTTGCGTCTG AAGATTTTTC AAAAGGAAAG TACGCACGTA CGCGATACAA GGTGATGAAA 6120

ATATGTGGGG CGTGCGCTTT TGTCCAGTTT CTATTGGATA CTGGTCGTAC CCATCAGATA 6180

CGTGTGCACG CGCGATACCT AGGATGTCCC GTTGTAGGAG ATCCGTTGTA TGGTTCCCGG 6240

AATATCTGTG GCATACCCAC AACACTCATG CTTCATGCGT ACGCAGTACG GTTTGTTCTT 6300

CCGAGAACGA AAAAACGCAT AACGCTGGTA GCGCCCATAC CGCTTCGTTT TGTTCGACTG 6360

ATACACCGAT TATCGGTTAG GTAGGGTGTG GCAGGTGCGT GCGTATATGC GTTTTACTTC 6420

AGCAGCTAAA TAGAAAGAAC CGATGGCAAG GAGCGCTTTA CGTTGCGCAA AACTGGCATA 6480

CAGTGCACGG GAAATAATTG ACGCAAAGTC TTCGCTCCAA AAAATTGGGA CTGTTTGGTG 6540

AAATGTCGTA CGAAACGCGT GGTATGTTTT TTGTATATCT GCATGTTTAG ATGTGCCCGG 6600

TATGGTTAAA AAAATTTCGC TGGCGGCATG TGAAAAAAGA GGGGGGAACT GGGACACTGC 6660

TTTGTCCGCC GCGCACGCAA AGAGTAAAAT GTATTGTGCA GAAGGTAGTA AAGAAGAGAA 6720

CGTACGACAT GCGCACCGTA TACTCTGcGT AgTGTGCGCA CCGTCAATCA CTATGaGTGG 6780

ATCTTCCTGc ATAATTTcAA AGCGTGcTGG cACGTATGCA CGGGaCAGTC CCCGCTCGAT 6840

TAACGTTTCG CTCACGGTAG GAAATAAATA TTTTGCCGCG CACGCAGCCA GTGCTGCATT 6900

TTTTGCCTGA ACAATATCGC ATAACGCGAG TGTGCAGTGT ATATTtCGAG CGAATAATCT 6960

GCCAACAGGA TGCGCGGCGT TAAAACTGAG AGTTGCaGTG TGTGTGAAGT GTTTTATtGA 7020

ACTTTCAATA TGTGTGACCA TATCtGGTAA GTAAAAGAAG GGAGCATGTT TTTCTCGCGC 7080

GATATGTTTA AAAACGTGCA ATGCATCTTC TGGCTGATCA AAACAAAAAA TAGGCGTATA 7140

GGGTTTGATA ATGCCGCCCT TTTCTTTTGC AATACTTTTT ATACGTGTTC CTAATATGCG 7200

CGTGTGTTCT TGTTCTATGG GGAGAAGGAG ACAGATACTA GGACAAATGA TGTTTGTTGC 7260

ATCTAGTCTT CCTCCAAGTC CTACTTCAAA AACGGACCAT TCCATGCGTT GTTGTGCAAA 7320

TAGCATGAAC GCCAGTAGCG TTATAAGCTC AAACCACGTC GCCTGGCCGT AGTCGCGCAG 7380

ATTCTCTGTT TTTTTCACCG TGTGGTATAC GTGTGTGCAC GCGCTTGCAT ACTCGGCAGG 7440

TGAAAAAAAC ACACCCGCGC GTGTTATTCT CTCTCTCGGA TCCATAACGT GAGGAGAAGC 7500

GTATAGCCCG GTGTTGAATC CAATTTCATT GAGTATCGCT GCAAgCATAC GTGCGCTGGA 7560

ACCTTTTCCC TTCGTGCCCG CAACATGGAT GCTCTGATAT GCGTTGTGTG GATTACAAAG 7620

CGCGCGTGCA AGTGCAGTCA TCCTGTGCAG AGGTGGTGTG CCGCTTGGGG GCATTTTCTC 7680

AAGCGTGCGA ATGCGCTCAA CCCAGGCGTA AAAATCTTGA AAAGAATGCA CCGGTATATG 7740

TGAAAnTCCG TGTGCGCTCG GTCGCACTAT AATATGCGGT ACGGAAGAGG CAGCAATCCT 7800

TGCCGGGAAA GAGAACTGAT GTACATTGCT AAGGTCTGAC ATTTGAGCTA AAATCCGCCC 7860

ATGAAGCGGG GGACTCTACC AAAAGATGTG TCAGGTATCA AGATTCACAT GATTGGTATC 7920

AAGGGCACTG GCATGTCTGC GCTTGCAGAG CTACTGTGTG CACGGGGTGC CCGTGTGTCA 7980

GGTAGTGATG TTGCAGATGT GTTTTACACG GATAGGATTC TCGCCCGTTT GGGTGTTCCC 8040

GTGCGTACTC CCTTTTCTTG CCAGAACCTT GCTGACGCTC CCGATGTGGT TATCCACTCT 8100

GCAGCCTATG TGCCTGAAGA AAACGACGAG TTGGCAGAGG CGTACCGGCG GGGTATTCCT 8160

ACCCTTACCT ACCCAGAAGC GCTGGGGGAC ATTTCCTGTG CGCGGTTTTC GTGTGGTATT 8220

GCAGGTGTTC ATGGAAAGAC GACCACGACC GCGATGATTG CTCAAATGGT AAAGGAGCTG 8280

CGCCTTGATG CGTCCGTCCT TGTGGGGAGC GCTGTTTCGG GAAACAATGA TTCTTGTGTG 8340

GTTCTTAACG GAGATACCTT TTTTATCGCA GAAACGTGCG AGTACCGTCG GCATTTCCTG 8400

CATTTTCATC CTCAAAAGAT TGTCCTCACC AGTGTTGAGC ACGATCACCA GGATTATTAC 8460

TCCTCGTACG AGGATATACT CGCGGCATAC TTTCATtACA TAGATAGGCT TCCTCAATTT 8520

GGTGAGTTAT TTTATTGCGT GGATGACCAG GGCGTGCGGG AGGTAGTGCa GCTTGCGTTT 8580

TTCAGTAGAC CGGACCTGGT GTATGTTCCT TATGGGGAAC GTGCGTGGGG CGATTATGGG 8640

GTCAGTATTC ACGGTGTTCA AGACCGGAAG ATAAGCTTCT CATTGCGGGG TTTTGCAGGT 8700

GAGTTTTATG TTGCGCTCCC CGGTGAGCaT AGTGTTTTGA ATGCAACCGG TGCGCTCGCA 8760

TTAGCACTGA GTTTAGTGAA GAAGCAGTAT GGAGAGGTTA CCGTTGAGCA CCTCAcGCTC 8820

TGCgGAAGGT ACTCGCTCTT TTTCAGGGAT GCCGGCGAAG GAGTGAAGTT CTTGGGGAAG 8880

TGCGCGGTAT TTTGTTCATG GACGATTATG GACATCATCC GACTGCAATT AAAAAGaCTC 8940

CGCGGGTTAA AAACGTTCTT TCCGGAAAGA AGAATTGTCG TCGATTTTAT GTCCCATACA 9000

TATTCGcgTA CCGCAGCCCT CCTCACCGAA TTTGCTGAGT CTTTTCAGGA TGCGGATGTA 9060

GTTATTTTGC ATGAGATTTA CGCCTCTGCT CGGGAAGTGT ATCAGGGCGA GGTGAACGGT 9120

GAACATCTTT TTGAATTAAC TAAACGGAAG CACCGGCGGG TGTATTATTA CGAGGCTGTC 9180

ATGCAGGCAG TGCCTTTTTT GCAGGCTGAA TTGAAAGAGG GCGACCTGTT CGTTACGCTC 9240

GGCGCTGGAG ACAATTGCAA ATTGGGTGAG GTGTTGTTCA ATTATTTTAA AGAGGAGGTG 9300

TAAAGTTCGG TTGCGGTTTC GCCAACATGG TGGTGGTGCC gGCTGTGGAT CTGGTGGATA 9360

TAGGGTGAAG TGAGACAGGC TGCGAATGGA TGGTGATGCA AAGAGCGGAG TGCGGAGGGG 9420

TGCGTGAGTA ATCGGTGCGA TGTGTCTGGA AATAAGGCGG TACgCATAGC AGTTTCAGGC 9480

GCGTCAGGGT GTGGTAATAC CACCGTGTCT GCATTGCTTG CGGAAAGACT GGGACTTCCC 9540

CTAGTGAATT ATACGTTTAG GAATATTGCC CGGGAGTTGG GTATCTCTCT TAGTGAGGTG 9600

CTCGAGCGTG CGCGGACGGA TAATCATTTT GATAAAGCAG TTGATGCGCG GCAGCTCTGT 9660

CTTGCGATGC GTTCTTCCTG CGTGGTAGGG TCGCGCCTGG CCATTTGGTT GGTGAAAGAT 9720

GCCGCGCTGA AGGTATATCT TTTGGCTTCA TTAAAAGAGC GGGTGAAACG TGTTCTCCAA 9780

AGGGAGGGAr GGGACGTACA GGATGTTGAG CGATTCACGT CTATGCGTGA CGCTGAAGAT 9840

ATGAGTCGCT ACAAAAAGTT GTATCGTATT GATAACACGA ATTACAGTTT TGCAGATCTT 9900

GTTCTAAACA CAGAAGGGTG CGATCAAGAA ACAGTGGTGA GTATTATTAT TGAAATGTTA 9960

CGCGCTAGAG GGATAGCTTG GTAGGGCTGA GCCAATCTGC GGGTGATATA GAAAAGTTTC 10020

AAAACGCCAT ATTGGATTTT TATGCACAGC AGGGCAGGGA TTTTCCGTGG AGAAGTACTT 10080

GCGACGCGTA TGnaTACTGG TGTCTGAGTT TATGTTACAA CAGACACAGA CGGAGCGGGT 10140

GTGTCCGAAG TATGCAGAAT GGCTTCATCG TTTTCCTTCT TTGGAGTCTC TTGCGTGCGC 10200

TCCATTTGCG CACGTGCTCC AAGCGTGGAT TGGATTAGGA TACAACAGGC GCGCTCGTTT 10260

TTTGCATCAG TCGGCAAAAC TCATTGTTGA AAGGTATTGT GCAGTAGTTC CTGATGACCC 10320

GAGTGAACTA AAGAAGCTCC CCGGTGTCGG TGACTATACT GCCGCTGCAG TTGCTTGCTT 10380

TGCGTACAAT AAGGCCACCG TGTTTTTAGA AACAAACATC CGTGCAGTGT TTATACGCTT 10440

TTTCTTTCCC GATACGCACC AGGTCAGTGA TCGGGAGTTG CTCTCGCTGG TCCGGTGCAC 10500

CCTGTATGAG GAAAATCCTC GGCGTTGGTA CTACGCACTG ATGGATTATG GGGCAGTTCT 10560

AAAAAGGAAG ATTACAAATC CTAATCGTCG CAGCAAGCAT TACGTGAAGC AGTCACCGTT 10620

TGAAGGTTCT CTGAGGCAGG TGCGTGGAGC GGTTTTAAGA GAGATAAGCG GCATGCAACA 10680

CGCGGTGCGC GAGAAAACGC TTTtCGCAAA GCTGTCCTTt GAGCACGAAA GATTGAGCCG 10740

CGCTCTAGAC TCGCTGGTAA GCGAGGGACT GGTAGTAAAA ACAGAGGCTG GGTATTCCAT 10800

CGCTGATTGA TTCTTTATGA CTCAAGACGC TTGAGTATTT CACAAATAAA GATGCCTTTC 10860

TCTTTTATTT CAATGGCATC TGATGTGAGG ATCAGCATGT TGGGCTGCTT TGCGTCAAGG 10920

CGAACGCGAT CGTTGGAGGT CTGTACCAGG CGCaGTATCT TTTTGAAGGC AGTGTCGCAG 10980

ACCTTGTCAA ACTCAATCAA TAAGCTTTCG TGTGTTTCTT TGAGTGAGAG AATGGCGAGC 11040

TTTTCGCACC GCACTTTGAT TTCTGCCACG GTAAACAACC CAGCTGCTTC CTCaGGGATA 11100

GGACCGAACC GGGTGATAGT TTCCGTGCGT ATGCGCTCAA GCTCCTCATG CGTATGAGCT 11160

GCAGCGATTT TTTTATACAG TTCCATTTTA ATTTCATCTG CGGCAATGTA CGTATGGGGG 11220

ATGAACCCTC GGTAATTAAG ATCGATGACG GTTTCTATCC TTTGCTCGTT TGGAGCATGT 11280

TGGAGGCGTT CTATTGCCTC TTCTAACAGC TGTACATACA GGTCGAATCC GACTGAATAG 11340

ATATCTCCTG aTTGTTCTTT GCCTAATAGA TTTCCTACCC CGCGAATCTC CATATCTTTT 11400

AAGGCGACTT TGAAACCCGC CCCAAGGTCA GTAAAGTCAG AGATCACCTG TAAACGTTTT 11460

ATTGCAAGGT CTGAAAGTGC CACGTCGTGA TAGTACAGCA GATACGCATA TGCTTTTTTG 11520

TCAGACCGAC CAACGCGTCC CCTGAGTTGG TAGAGCTGGG AAACCCCGTA CATATCAGCT 11580

CTATCTATGA TGATAGTATT TGCATTGGGA ACGTCGATAC CATTTTCAAT AATGGTGGTA 11640

GAAAGCAGGA GCTGGAACGT TTTTTGATAA AACCTTTCAA AAATGTCTTC CAGTTCTTCT 11700

GACCCCATGA GACTGTGGGC AACGCATATG GATAGCTCAG GCACGAGTTT TTGGAGCATA 11760

CACTTTACGG ATTCTAAGTT TTCGATTCTG TTATGTAGGT AAAAAATCTG CCCCTCACGA 11820

TCTAGCTCTT TTCTGATTGC AGTGGCAACA AGGTTTGGAT CAAACTGCTG GATAACCGTT 11880

TCTATAGGTA GGCGGCCTTC AGGAGGGGTG GTGAGCAAGC TCATGTCTCT GATTTTGAGC 11940

ATACCCATGT GAAGCGTTCG GGGAATGGGC GTTGCACTGA GGGAGAGACA ATCTACATTA 12000

GTTTTCATCT GCTTTAATTT TTCTTTATCC TGCACACCGA AACGTTGTTC CTCATCGAGG 12060

ATCATCAACC CAAGATCCTT GAAGGACACG TCCTTTTGGA TAAGCCGGTG GGTACCCACA 12120

ATAAGATCGA TATCTCCATG CGCGAGTTTG GCGAGTATGT CCTTTTGTTC AGATTTAGGA 12180

ACAAAGCGTG AGAGCTTCTC GATTCTGACG GGAAAGTGTT TAAACCGATT GCAGATTGTG 12240

CGAAAGTGTT GTTCCACTAG TAAGGTGGTA GGGGTGAGGA ACACCACTTG TTTTCCTCCC 12300

ATTACCGCCT TAAATGCCGC GCGCATTGCA ATCTCTGTTT TTCCGTATCC GACATCTCCG 12360

CACACCAGCC GATCCATGGG GACGGCTTCT TGCATATCCT GTTTGACTTC TTCAATGCAT 12420

ATGCGCTGAT CGTCTGTTTC TTCGTAGGGG AATGCTGCTT CAAACGCATA CTGCCATTCG 12480

TCATCTTTTG GGAAGGCGTG GCCGCGCGTA GTTTTTCGCA GAGAGTAGAG TTCCACTAGT 12540

TTTTGCGCGA TGTTTTCAAC AGATTTTTTG ACACGTGCTT TTCTCGTTTC CCATGACTTT 12600

GACCCAAGGC TATCTAAGTG AGGTTTGTTC CCTTCATTTC CAATGTAACG TTGCACCAGA 12660

TGTGCCTGCT CAATAGGGAT AAGGATCGTT TCTTCCTGTG CATAGAGGAG GTTTACGTAA 12720

TCACGTTCTG ACTGTGCTGT TTTTATGCGC TCTATTCCCT TAAATAAACC GATGCCGTAC 12780

TGCGCATGCA CCACGTAATC CCCGGGATTT AATTCCACAA ATGTGTCGAT AGGCGTGCTC 12840

CGTGCGCGTT GCACTGATTG AGGAGTTTTT CTGCGGCGAC CGAAGATTTC GCCTTCTTGA 12900

ACGATCAGTA TTTTGAGAGC AGGAATGCTA AATCCTGCAG AAAGCGCGCA AGGTAGCACA 12960

GTGACGTCGC AACCTTTGAC TAGTGCTCTG ATGCGmrTGC CTGCTGCTCA CTTTCTGCAA 13020

AGACGAAAAC GTGCCATCCG TCTTTTGAAA GACGGAGTAG CTCTTCTTTG AAGTAAGGAA 13080

TGTTACCGAA GAAGCTGCGT GCAGGATCGC TTGCCAAGCA TATACTTTCG CACGCTGGCA 13140

GCTGTGGAAA AAAGTGAGTG AAATACACCG TGTGCAGGTG GAGCGCGCAG ACAGCGGAAA 13200

AATCGAGCAC TATGTGTTCT GGTTGAGGAT ACCAGCGCGC AGtACATGTT CGTGCGCGAG 13260

TTGCATTTTA TGGTAGAGGT TCCGACACTC GTCTTGGAGC GCGCGTGCAC CGTTGTGCTG 13320

GCGTTCGTAG TCAAGATAAA AGACGCTTGG GGGTGAAGGG CTGTGGCGAA AATATTCGAG 13380

AACGCAGgTG GGACGTTCAA AGCACAGTGG ATAGAACATT TCCTCCCCTT CATACGTTTT 13440

TCTGTGGGTG AGTTCTTCGA TACACGGGAC GCAGTGGGCA GGACATTCGG ACAGTTTTTG 13500

GAGATTTTGG TGGAGGAACG CTATACGCTC CTCACTCCAA AGAATTTCTT TTGCAGCGTA 13560

CAGTGTGCAC GCAGATACCT CTTGCAGGAC GGCACACGTG GACACCGCCA GTATATGGAT 13620

ACGTTCTATG GTGTTAAAAT CACACACGAT TCGGTACGCT TGTGTGTTGT CAGCAGCCTG 13680

CGCTGCGGCA GCGATATCGA GAATTTCTCC CCGGAGAGAA AACTCTGCGC AAgcGcTGaC 13740

GTGGTCGACA CGTGCATATC CCCATTGCAT AAGCTGGGCA GCGAGCGTGT GGATCTCGAT 13800

GTGCTCTCCC ACACGGAAGG AGCGTTTGAG GGTACGCACA TAATCGAGGG GAGGAACGGG 13860

GGTGAGCAGT GCACGCTGGG TGAAAACGaA CrcGCATGgC gGTaTGCATC GcGCTGTGCG 13920

AGTGCGCACA nnCTnCTnAC CCGGTGAGAG AACACGTGTG CGTTAGGTGA GACAGGGCGG 13980

TAGGGCAGCG ACCCCCACCA GGGGCAGCAC GCGCGTAGGA ACTGCTGCAT GTGCAAGGTC 14040

GGTGCAGACG GCGGCGACGT CCTGTTTCGG TAnGGACTAC GAGCACTATG TGTGCGCAAC 14100

ACGTGCGCAC GTTATTCGCC AAAAAAGTAG GACCGCAGTC CAACGGTGCA CCCTTTCAAA 14160

CGCGGTAGGG AAAAGCGTGC GGCACCAGCG AnnGCAGCAA TTGCTGGAAG CTCATTTCCA 14220

AGAAATGGAG TATGCCACGC AACA 14244

(2) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 2109 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

AATCGACGAA AGGTATTGTA GCAGAGAGAA TACTCAAGCC ATGCGTAAGG AGAAAAGTAA 60

ATGGCAAGTT TAGATCTACC TAAGAGTCCC AATGTGTTTC ATCCCGAAAA GCCGAGTGCG 120

GTTGGGTCAA GGAATTCACT GGCGCAGGAC TGTCGTGACC AGCAGCAGGA GGTGAACCAG 180

CTAATAGAGG AAGAGACAAA CAAGATTCTG CACCACCTGA ACACTAAACT GCCGAAGaGG 240

TTCTCGAGCG TCTGGACGTA ATGGGTGGGT TGAAGGAAAA GTTGTATAAC TACTTCAACC 300

AGAATTACCA GAACATGTTC AACCGGTACA TGGTGACTGC GGAAGACGAA ATGCTGAAGA 360

AGGTCCGTGG TTTCATCGAC CGAGAGGAAA TGAAGGTGTT GAACCGTTAC ACGCCGAAGG 420

AGATTGCCAT CCTACTGGAT GAGGTTGCGG GAGCGGATAA GTTCAACACC GGAGAGATCG 480

AGAAATCGAT GGTGAATATG TACGGGCACT TGCAGGGTCA TATACAGCGG GGTGTGAATG 540

AGCTTGAGAC GCACACCAAT TCTTTGCTGC GTCAGAAGGT TGATGTGGGT GCTTTTGTCC 600

GCGGAGAGAA TGCGTATGCG GTAGTCAAGT GTGCGTTCAA GGACAATCTT GCGCGTCCTA 660

AGACCGTCAC TGACGTGAAG TTGTCTATCA ATATTCTGGA CTCAGAGTTA GTTAGCCCTA 720

TCTTCCATTA CCAGACGACG GTAGCGTACC TTATTAAGGA TCTCATCTCC AATCACTACA 780

TAGATGCCAT CGACAAAGAA ATTGATCGCG TGAAGGACGA GCTTATCGAC CAGGGTAAGG 840

AAGAGATGTC TGATAGCAGT ATCATCTTCG AAAAGATGAA GATGGTGAGC GATTTCACCG 900

ACGATGACTG CGAGAAmCCT GACAGCAAGC GCTACGAGCT TATTTCGCGG GAGTTGATGG 960

AAAGAATCAG CAATTTGCGC GCGGAAATTG ATCCGGAAAC TTTCGACCAA TTGAATGTTC 1020

GCGAGAATAT CAAAAAAATC GTTGACCTTG AGAACATAAG GAATCGTGGC TTTAACACGG 1080

CTATCAATTC GATTACATCT ATCCTTGATA CGTCGAGGAT GGGGTACCAG TATATCGAGA 1140

ACTTCAAGAA TGCGCGCGAg CTTATCCTTC GTGAGTATGA TGACACAGAT ATTTCGAATC 1200

TTCCTGaTGA GCGTTACCAG TTGCGCTTAA AGTACCTCGA TAATGCTCAG TTGATTGAGG 1260

AGCGTAAGGG GTATGAGGTG ATGCTTCGTT CTTTTGAGAC GGAGGTGGAT CATCTATGGG 1320

ATGTGCTGCG TACTAAGTAC GATAAGTCTA AGGCGTCTAG GTTCATGGCG AAGATTACCG 1380

ACTTTGATGA CCTTGCTAAG GTGTACAAGA AGCATATAAA GAAGCATTAC AAGGATAAGA 1440

CTGGTGAGCC CGTGTACGAG GATATTGCGA AGGTATGGGA CGAGATTGCT TTTGTGAAGC 1500

CTGCTGAGAC CGAGGTGGAG CGGATGAATC GTACGTTTGT GTACGAGAAA GACAAGATGC 1560

GAAGGAAGCT TATTCTGATG CGTGGGAAGT TAAAGGGTAT GTATGATTAC CAGTATCCTA 1620

TTGAGCGTCG GGTTATGGAG GAGCGTCTCG CGTTCTTGGA ATCCGAGTTT AACCGTTTCG 1680

ATTACTTGGT GAATCCTTTT CACTTGCAGC CGGGCTTACT GCTCGATATC GACATCACGT 1740

CTATAAAGCG CAAGAAGGCG ACGCTCGACG GTATGGCTAA CGTGCTTAAT GAGTTCTTGC 1800

ATGGTATCTC TAAAGGATTT GCGGACGCTG CCTTTGCTTC GTTTAGTCGT CGTCGTTCAA 1860

CGGTGCGTGC TGATATCGGT CAGAGTTTTG CTAGTGACGG CAgTGCCGAC CAGAAGGAGT 1920

CCAGCGGTAG GGTGGCTTTT ATGGATATGG TAAATGAGAC TCCTGCGCTT GAGTCTTCCG 1980

TGGCCGCTGA GCAGGTGGAT GTGCGCTCGG ATGTTGGAAT GAAGACGAGA AAGGTGGCGC 2040

GGTGGATGCA GGCAAGGGTC GACGTGGTAG ACGGTCTGCC ATTCGCGAAt CTAGCGAGAT 2100

TGTAGATAC 2109

(2) INFORMATION FOR SEQ ID NO: 4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 9848 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

CTGGACATGT TTCCTGCCTC TGAACCTTGG GTGAGGGAGT TTGCACAGAG GGTGGGGATT 60

CACGTGCAAG AAGGTGCACG GCTCGTGAAT TTGCCTCGTC ACCCTAGCCA AATCTATGAA 120

GCTTTTTTTG AAGGAATTGT GCTGTGGTGT ATTTTGTGGT GTGCGCGTCG GGTAAAAACG 180

TATAACGGCT TTTTGGTGTG TTTGTATGTG GTGGGGTACG GAGTGTTTCG TTTTTTTATT 240

GAGTATTTTC GTCAGCCTGA TGCGCATTTG GGGTACAGGT TTTCCGCCAC GCAATCGTCT 300

CCGATTTACC TTTTCCAGTC ATGGAGTGAT GTTTCCACCG GGCAGATTCT GTGTGTTCTA 360

ATGATTCTCG CAGGTTTGGG TGGGATGTTC GCACTTTCGG CGTATCACAA GCGGGATAGT 420

GTGCGGAAAG CGCGTGTATG AAAATGAAAA GAATGCACCG ACTGGTCCAT CAGCCGAGAT 480

GGGGTGGCGC GATGTACCTT GCGTATAATG CGCAAAGGTG AGCCGATTcT TCTTCCGTCG 540

TCCTAATGAG CTGGTCTTCT TCTGGTGTAC CAAGCACCGC ATTGCGCTCC AGAGGGTTTC 600

GGTATGTGAT TTCTCGAAAG GGGTTCAGAT AGTCTTTGCA TCTCGTTTTG TGAAtTTGTG 660

TTACTCTGTA CTTTTCGATG GTAATTGCGT GTTGCACGTA GGTGCAAAGT AGCGGTGCGT 720

CTATCCTTGT TCGTGCAAGA GAACATAACA CTGCATAGGC GTGGAGTGCT TTATCCGTTT 780

CGTACGTGGC AGCAAGCGCG TCGTATTCTT TGTGGATGTG GTGCCAACTT TTGAACTTTC 840

CCGTTTCAAT GTTTTGTAGA AGTGTCTCAA ACTTATCTTC AGGTACGAGT TGTCCTCCCA 900

TGTTTACCCA GTTGTGTGTG ATAGTTTTAG GATCGTGCGA AGTAAAGTGG GAGATGCAGC 960

GTGTTGTTTG CTCAAAGAAA GACAGGAGCG TTTTTGTGCA GTACCATATG AGTATGTCTC 1020

GGTATGCTTT GCGGGCTTCT AAGGGTTTTA GCACAAGTGT TGATCGGGTG GAGTGTTCTA 1080

TGCCGGTGAG CAGTACGGGG ATTGTCTGTG CTTTGGTGGG ATGTTTGAGG ACAATGTCTT 1140

CGGCAGTGAG CGCGCTGTTT CCCGCGTTAA CCCACGCGCG CTCTATCGCG CTGTCTAGTA 1200

AAGCGAGTGC GTTTTCAATT TCTCCTATTG TGTCGGGAGC GAAGACAGAG ATTTCTACTG 1260

TTTGTGTTTT CGTTTTTCGT TTGTCTCGCG CAGAAtTTTT TTTCGTTGCG TTCAATCGCA 1320

TACATGTTGT AGAGCCAGTA GTACGCGGGC ATGATTTCTA ATCTGTTTTC CCGTTCATTG 1380

TTGGTGACAA GACTAAAAGG AAATGGAATA TCAAGTTCTG CGGGGTAATT TTTCCCCGTG 1440

ATTAATACGA AAGAAGCAAA ACGACAATTA TGTTTGAGGG TACTTGCAAG TCCTGACCAG 1500

AACCCTCGTC CCGCAATAAT TTCTCCGTCA TTTTTGCGTG TATTGTGATT TGACCCAATG 1560

GTGGAGCCAG CAGCAATATT TGACTGACCA CGTATGAGTG CTGCGATGAG GAAAGAGTTG 1620

TTGTGGTGTT GTTCATGGTA GGGAAAGATG AGTGCGTTAA GAACCTCACA ACACGAAATC 1680

GTGGAGTTAT CACCTAAAAC TGAGTGAATG AGTCTGGTAC CATATTTAAG TGCGCAgTTA 1740

TTTCCCAGTA CAAAGCGCAC CGCCTTTACC CCATAGAACA CACGGCAACC ATATCCGATC 1800

ACCCCGTTCA CTAACTCTAC TCCTTCTCCT ATTTGCGTAG GTTCCTGGAG AGATGACTGA 1860

ACAGTAAGGT TTTTTAGTTT GTTTGCTCCT TTTACGTATG ATCCAGGACC GAAGcACACA 1920

TCCTTGATGA TACGGCaGCT TTTGATAACG GATTGTGTTT CaATCGTTCC gTAGTATCCg 1980

CgGCGAGTGT CGTGTTGTTG TTGAGTCATT GACTCGAAGC GTTGCATGAG CAATGTTCTG 2040

TCTCGGTGGC ATGCCCACAA AAACGCGTCT GCTGCGATCA TCCCTACGAA GGGAAATATT 2100

TTCCGTCCGC CTGTTTCGTT GAGAGGATCA ATGGTGATGC GGACTGCTTC TTGTTCTCCG 2160

TCTTTTATAA TTCCTGCGCC GAATTTTGCG TGGTTAGTGG TACACAGCTC ATCAATCCTG 2220

CTGAGGATGA CATGATTTCC AATTATGTAG TGAGAAATAT ATGCGCAGTG ATGGATGGCG 2280

CAATTTTCTC CGACGTCGCA CGAAaTGAGT GTACTGTGGG TAATACCGGT TGGTACGGTA 2340

AAGTCGTGAT ATCGCAGAAA GgCTCGCTCG AGCGwasCGA TGCGTACGAG CCCTGCAAAT 2400

GATGAATTAC GTATGAGTGA CGCGTCGAAC GGATCTGCTA CTAAAACGTC GTGCCAGGTA 2460

TCGCAGTGAT TGCCCTTTTG TATAAGGGTG TGAATTTCCT CCTTAGACAA TGGTCTCCAT 2520

GCACGTGGGG GTTCGGCGCT CTGGGAAAAG CGGAGATAGT ACTCGTCTTT TCCGCGAGGG 2580

ATATGGGTTT GTGTGATGAA GTGGTATCCA AAAGAGGGTA GGTCTAAAAT TTGCACACGC 2640

ATTCTCCCTT TTGGATGCCC ACTATAGGTG GTGAATTTTT ACATGTAAAT AAGGAATTGG 2700

GGTTGTGATG GGGATGGTGA TTTCCTGCAT GTTTACTTGA CATGACATAT TAGGAATGGC 2760

TAGATTGGGG CCCaGTCTTG TTTTTTAGCG TGCATTAGAA GTGGATGTaC TGGGGAGGAT 2820

CgTTGGCGGA TAACAAAAGC TTGCGGATTA ATGGAAGTAT TCGGGTACGA GAAGTGAGGT 2880

TGGTTGACGC TGTAGGGCAG CAGTGTkGGG TGGTGCCCAC CCCTGAGGCG CTGAGAATGG 2940

CACGGGATAT CAATCTTGAT TTAGTAGAGG TcgcTCCGCA GgCGAGTCCG CCGGTGTGCA 3000

AGATCCTGGA CTATGGGAAG TATCGCTTTG AGATGGGCAA AAAGTTGCGT GACTCGAAAA 3060

AGCGACAGAG ATTGCAGACG CTCAAGGAGG TGCGTATGCA ACCGAAGATC AACGACCATG 3120

ACATGGCGTT TAAGGCCAAG CATATACAGC GGTTTCTCGA TGAAGGGGAT AAGGTGAAAG 3180

TGACTATCCG CTTTCGTGGA AGGGAGCTTG CGCATACCGA TCTGGGTTTT AACGTGTTAC 3240

AGAATGTGCT TGGCCGTCTG GTGTGTGGGT ATAGTGTTGA GAAGCAGGCA GCAATGGAAG 3300

GTCGGTCTAT GTCCATGACG CTCACTCCGA AGTCAAAGAA ATGATGGAGT GTCGGGTAAC 3360

TGCAGTTCGT GTTGTTGGAT AAAGGGGAGA AAGTATATGG CTAAGATGAA AACGAAAAGC 3420

GCAcAGCAAA GCGTTTTAGT GTAACCGGGG CTGGTAAGGT AAAGTTCAAA AAGATGAACC 3480

TGCGTCACAT TTTGACGAAA AAGGCCCCGA AACGCAAAAG GAAATTACGT CATGCGGGTT 3540

TTCTGTCAAA AGTTGAGCTT AAAGTGGTGA AGCGGAAGCT GTTGCCTTAC GCGTAGgTGG 3600

CAAGCGTGAG AGGACGGAGG AGCGTGGTAT GTCTCGATCG TTGAGTAGTA ACGGCAGAGT 3660

GCGCCGGAGA AAGAGGATTT TAAAGTTAGC CAAGGGCTTT CGGGGTAGGT GTGGCACGAA 3720

TTACAAGGCG GCGAAGGATG CGGTCTCGAA GGCTCTTGCG CATAGCTATG TTGCGCGGAG 3780

GGATAGGAAG GGGAGTATGC GCAGtTGTGG ATCAGTCGCA TCAATGCATC GGTTCGTACG 3840

CAGGGtTGAG CTATTCTCGC TTTATGAATG GTCTCTTGCA GGCTGGGATT GCGCTTAATC 3900

GCAAGGTTCT CTCCAATATG GCAATTGAGG ATCCAGGTGC GTTTCAGACG GTGATCGATG 3960

CTTCTAAGAA AGCTTTGGGG GGTGGAGCGT GCTAAACCTC GGTCAGGTAA AAGTGCTGGA 4020

GGAGAAGGTT GCGAAGgCGG TGCACCTTGT CCAAATGTTG AAGGAAGAAA ATGCCGcgTT 4080

GCGGgCTGAA ATTGATGGAC GTGGTAAGCG TATTACGGAG CTGGAGCAGC TGGTGCTTGs 4140

CTTTCAGGAT GATCAGACGA AGATAGAGGA AGGAATTCTT AAGGCACTGA ACCACCTGAG 4200

TACATTTGAG GATTCTGcGT ATGGAGAAGC GCTTACGCAA CACGCGGCGA AGgTTCTAGA 4260

AAACCGGGAG CATGCGGGGC TGTCTGAAGA ACTTACCAGC CGTACCCAGA TGGAAATTTT 4320

TTAGTGGTCA GTGTAAAGGG GCAgTTGCAC ATCGATCTGT TGGGAGCGTC TTTTTCCATC 4380

CAGGCTGACG AGGACTCCTC GTATCTGCGT GTGTTGTATG AgCATTACAA GATGGTGGTG 4440

TTGCaGGTGG AGAAGACGTC aGGGGTCCGC GATCCcTTAA AGGTcGCGGT GATTGCgGGT 4500

GTGCTTCTCG CGGATGAACT GCATAAAGAG AAGAGGAGAC GTCTTGTACA GTCCGAGGAA 4560

GATCTGCTGG AAATAGGGGA GTCTAnCCGA GCGTATGCTC GAATCCATCA GCAAAGTGGT 4620

GGACGAGGGG TTTGTGTGCG GGCGCGATTG AGGGTTGTGT CCTTCTTTGT GTACGGGACG 4680

TCCTGCGGTG ACGCTGTGGG TGGACGCGGA CTCATGCCCC GCGCgcGTCC GCGTACTTGT 4740

CGCGAGAGCG GCAGCGCGCC TGGGGTGTGT GGCTCGATTT GTGGCCAACC GTCCTATCCC 4800

TCTCGTGCAA AGCCCGCATT GTATCATGGT CGAGACTCAA CCTGTTGACC AGGCTGCGGA 4860

CCGTCACATg CATCGCGTAT GCGCGAGCGG GTGATTTGGT CGTCACGCGT GATATCGTGC 4920

TTGCAAAGGC AATTGTAGAC GCGCGCATCT CTGTTATCAA CGACCGGGGT GATGTGTATA 4980

CGGAGGAGAA CATACGCGAG CGACTCTCGG TGCGTAACTT CATGTACGAC tGCGAGGGCA 5040

GGGACTCGCC CCTGAAACAA CGTCACCGTT CGGCAGGAGG GATGCCGCAC GCTTCGCAGA 5100

CTCCCTAGAT AGGGAAACCG CGAAgcTCCT GCGGCTTGCC AGGCGGCGGG AGGCGAAGAC 5160

AGGGGAGGAG CAGTGCGACT GGCCCTCCGC GCAAGGGAAA AGCCAAACCG GCCGCCGGTG 5220

ACCGCACGCA AGACACTAAG AGTCCAAGGC CGGGCGGGTG GACTCCTAGT GTCTTCTACC 5280

GCTTCTGCGA GATGAACTTA AGCAAATCCA CCACACGGTT TGAGTAACCC CACTCGTTGT 5340

CATACCAGGA CACTACCTTG AAGAAGCGCT TCTCGTTCGG GAGGTTGTTC TGCAGCGTCG 5400

CCCTGCTGTC GTAGATGGAG GAGTACTGGT TGTGGATGAC GTCCGCGGAT ACAATATCCT 5460

CGTCGCAATA CTGCAGGACA CCCCGCAGAT AGGACTCCGA CGCCTTCTTG AGCATCGCGT 5520

TGAGGTCCGC AACGCTCGTC TCTTTTTCCG TGCGGAAGGT TAGATCCACC ACGGAACCGG 5580

TTGGTGTCGG GACACGGAAG GCCATCCCCG TCAACTTACC TCTCGTAGAC GGCAGCACTT 5640

CGCCTACCGC TTTCGCAGCT CCAGTGGTGG AAGGGATAAT GTTAACCGCT GCAGCGCGGC 5700

CTCCGCgCCA GTCCTTCAAA GAAACCCCAT CTACAGTTTT TTGCGTTGCG GTATAGGAGT 5760

GGATAGTCGT CATCAGTCCC GTTTCAATAC CGACTCCCTC TTTGAGAAAG ACGTGCACTA 5820

CCGGCGCGAG ACAGTTGGTA GTGCAGCTCG CGTTGGAGAC GACCTTGTGC TCAGCAGGAT 5880

CGAACTCATG CTCGTTCACC CCCATTACAA TAGTCTTCAC CGGCTTAGAC GCATCCGAGC 5940

TCTTAGCCGG AGCACTGATG ATGACTCGCT TTGCTCCTGC TTCAAGGTGA CCGTATGAAG 6000

ACTCATTCGC GTAAATGCCG GTGGscTCAA TAACCACCTC AATACCAAGA TCCTTCCAGG 6060

GaAGTTGGGA AGGCTTTAAG CCGCGACCGC AGACACACTT GATCCGATGC CCGCCCACCT 6120

CGAGGATATC CTCGGCAGGA GCACTGAGAC TAGAACCCAT TTTGCCCTGC ACGGAGTCAT 6180

ACTTTAGCTG ATAGGCAAAG TAGCGCGCAT CGGTGGAAAG GTCTACAACT GCCGCCACGT 6240

CGAACTCTTT CCCCAACAGC TTCTGtCCGC CATGGCCTGG AGTACGAGAC GCCCGATACG 6300

CCCAAAACCA TTGATTGCAA CTCTCATTTG CCCAACCTCC TCTAAAAAGA GCACACATCC 6360

CGCGCAACGC TATCTGAAAA AAGATCGGCA CGTCAATCCC TCTTTGCTGT AGGGCTCCCT 6420

TGCATTTTTC TATGTGCCCA GATACCATGG CCTCGCCTTG GAAGGTCTGG CCTCTAGTGG 6480

AAGATTATTA CCGCGTGCTT GGTGTGTCGC ACCGTGCCTC GACCCCTGAA ATTAAGTGTG 6540

CCTTCAGAAA GAAGGCAAAG GCGTTACATC CGGATCTCGT TTCCCATACT GCAGAACTTG 6600

AGTGCGAGGC GGTAgCgCGC GAGCGCgCTC TTCGCCGTAT ACTCACCGCA TACGAGGTGC 6660

TCTCTGATCC GGGGCGTCGC GCGAAATTTG ACCTCCTCTA CGCGCGTTTC TGCGCACGTC 6720

CTGCTCCAGC GGGCTTTGAC TACCGCGTGT AmCTGCGTGC GCAGGtACGC TCTGCGCGAT 6780

GGTGGAGCTT ATCTTGTTTG ATCTCTTTCA CGGTTTTGAG TGTGACGCTG TCCGCGCGTA 6840

CTTGTCCCTC AAGTGTCGGC CAGAAGGGTT CAACCTCGCC ACTCACCTTA CACGAGAGGA 6900

TTTTATGGAC TGTGGCTTTG TGCTCGCAGA GGAATTGCAT GTACGGGGAG AGTGCTATGA 6960

ATGCTTTACT TTGCTCCAGG ACATCGTTTT TGAAGAATTG CGGTGCGCGT ATTTTCGTCA 7020

TTTTTTTCCT GAAGTACTGA AGCTCGCTGA GCATATCGCG CTCGGTAcTG CGTCTGTGCG 7080

TGGTCGCAAC GGTAAATCCT GCGTATACTG CGCGCGCGCC ATGCCTGCTT GCCTGCGCAA 7140

GAAATTGTCA CCTTCTACGC GTGCTTAGTT GAGTATTACG AACGTACGGG AGACCGCAGc 7200

GTGCGCGTGG CTATGCCCAG AAGATGGATT CTGTCAGGTG AATGTTTGAC TGCACCCTGG 7260

CGGAGGAGTA CCGTGGTCCT GGGGGACCTC CGAAGGcTGG AGGTCCCCCT GCAGCTAGTG 7320

AACGGACAGA GGAGGGACGC TTGAGCAGGA AGGAAAGGAC CTCATGATCC GCATTAAAAC 7380

ACCAGAACAA ATCGACGGTA TCCGTGCCTC TTGCAAGGCA TTGGCGCGCC TTTTCGACGT 7440

TCTTATTCCG CTTGTCAAAC CGGGCGTTCA AACCCAGGAG CTTGATGCGT TTTGCCAACG 7500

CTTCATCCGC TCAGTCGGTG GTGTTCCTGC CTGGTTCTCG GAAGGTTTTC CTGCCGCTGC 7560

TTGCATTTCA ATCAACGAAG AGGTCATCCA TGGTTTACCT TCAGCGCGTG TGATTCAGGA 7620

CGGGGATCTT GTTTCCCTTG ATGTTGGTAT CAACCTCAAT GGATACATTT CTGACGCGTG 7680

TCGTACTGTT CCTGTCGGTG GAGTTGCACA CGAGCGACTA GAACTTTTGC GTGTAACCAC 7740

TGAGTGCCTC CGTGCGGGCA TTAAAGCGTG CCGTGCCgGA gCnCGyGCGC GCtgTTTCTC 7800

GCGCTGTATA CGCTGTTGCA GCACGGCACC GCTTTGGCGT GGTGTACGAA TATTGCGGAC 7860

ATGGCGTGGG GCTTGCCGTG CATGAGGAGC CGAACATCCC CAATGTGCCT GGCTTGGAAG 7920

GGCCTAATCC ACGTTTTTTG CCCGGTATGG TAGTCGCGAT AGAACCCATG TTGACGCTTG 7980

GCACAGACGA GGTGCGCACC AGTGCAGATG GCTGGACGGT GGTAACGGCA GACGGATCGT 8040

GTGCCTGCCA TGTGGAGCAC ACTGTGGCAG TTTTTGCAGA CCACACGGAG GTTTTAACAG 8100

AACtACGGAA GTAGAGCGTA CCGGCTAGTC AGCTATCTTA AGTGTGCGCG GTGTGCTGAT 8160

AGTACATGCA GGGAGCAGTT TGTGCACGGT AGGCAGCGTG TAAGTGTACG TGGCGGGCAC 8220

AGGTGAAGAG GGGATAAACT CGTAACCATA TCGCTGTGTG CTGCTTTTAA CCCGGGCTGT 8280

GTCGGTAGGG GTTTGGGTAC GCGCAgGGAC GTGGAGGGAC TCATGAACAT ATTGTTTACC 8340

TCGTTTGTGT GTGGGGTACA TGCGGTATGC CGCAGTTTTT TTACAGCAGC GGCGTTGCTC 8400

GTTTTTATCT GCTGCTCTGG TCATCCAAGT TCTGCGCGTG TGCCCtCTGC AGACACGATA 8460

GCTCGGCGCG TTGCCGGAGA CAGTGGGAAC gCTGGGGGGC GGACATTACT TCCTGTGGGG 8520

GTTTCgCGTG AATCGGTGCA GCTGTTAGAA CGGCTGCAAA ACGCGAACCG TCAGGTAACT 8580

GCCGAAGTGC TGCCTTCAGT AGTGACGCTG GATGTGGTGG AGACCAGAAA GGTTCGGGTA 8640

CGTGATCCGT TTGGCGGTTT TCCGTGGTTT TTCTTTCGTG GTCCTGAAGG TCCGGGTGCG 8700

GGGnCTGGCG GTGGTTCTGG AAACAAAGGG GAAGCTGAGG AACGGGAGTA CAAAACGGAG 8760

GGACTTGGTT CTGGAGTCAT TGTAAAGAAG ACAGGGAAGA CGCATTACGT GCTTACCAAC 8820

TATCACGTGG CGGGTAAGGC TAATGAGATA GAGATTAAAC TGCACGATGG CAGAATCGTA 8880

AAAGGTAAAC TTGTCGGTGG TGACCAGCGC AAGGACATCG CGCTGGTCTC CTTTGAGGAC 8940

GCAGACCCAA ATATCCGTGT TGCCGTCCTT GGTGACTCGG ATGCAGTACG GGTAGGAGAC 9000

ATTGTGTTCG CAGTTGGCTC TCCTCTTGGG TACACTTCCA CTGTAACGCA GGGGATTATC 9060

AGTGCGCTGG GTCGCTTTGG GGGACCGGGC AACAATATTA ATGATTTTAT TCAAACAGAT 9120

GCGGCCATAA ACCAGGGCAA TTCCGGGGGA CCAATGGTCA ATATTTATGG CGAAGTGATT 9180

GGGATTAACG CGTGGATTGC CTCCTCAAGT GGGGGATCGC AAGGGATTGG TTTTTCAATT 9240

CCTATCAATA ATGTGAAGTC GGATATCGAA TCATTTATCC AGTACGGGCA GGTGAAGTAC 9300

GGGTGGTTAG GCGTGCAGCT GGTGGCAACG GATGCGGACA CCGTAGCATC GCTTGGTATT 9360

GCAAAGGGTA CAAAAGGGGT GCTTGCGGCG GAAATTTTCT TAGGTTCTCC TGCGCACAAG 9420

GGGGGACTGA AACCGGGCGA TTACTGTGTA AAACTGAACG GAAAAGAAGT AAAGGATGTA 9480

AATCAGTTTG TGCGGGATGT CGGCGCGCTG CGCATTGGGC AAACAGCAGT ATTCGATTTA 9540

ATTCGCGGTG GTGTGCCGAT GACGCTTTCG GTGCGCATTA CGGAGCGTGA TGAAAAAATA 9600

GTAAATGACT ACTCAAAGCT TTGGCCTGGG TTCATCCCAC TGCCGCTTAC GGAGGCCGTG 9660

CGTAAACGTT TGGATTTGAA AGCGTCGGTG CGTGGTGTGC TAGTTAGCAA CGCGCAGAGC 9720

AAAAGCCCTG nCGGCGCTGA TGGGATTGAA GTCGGCGGAC ATAGTAGTGG CGGTCAATGA 9780

TCAAAGAGTC TCGAGCGTGC GTGAGTTTTA CGCGGnGCTT GCACGTCAGA CGAGGGAAGG 9840

TGTGGnTT 9848

(2) INFORMATION FOR SEQ ID NO: 5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 7415 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

CAAAAGGAGA TGTCATTCTA TGTGGAGCAT GCCATTCGCG TATGCACCTG GGTTGTGCGC 60

AGACACTTCC GCCTGATGTG CGCACCAGAT AGAGTATGTT GTGGAACGCA CTCGGTCTCA 120

TCGGGGAACT ACTGTTGCGC TTGCTATAAA TTATGGGGGA AAAGATGAAA TTTTACGTGC 180

GGTAAAAAAG GTTTTGTGCA GCACTTCGTG CCCGGATGGT GAGCTTCTCA CCGAAGAAGC 240

TTTCGGCGCG TGCCTTGATG CGCCGCAGTT GCCGAGTGTC GACTTTCTCA TCAGAACAGG 300

GGGTCAGCAA CGCATGAGTA ATTTTTTGCT TTGGCAAAGC GCGTACGcGG AGTTCTATTT 360

TACCGATATC CTGTGGCCTG ACTTTCGGGT AGAAGACATG cTGCGCGCCC TGGATGAGTA 420

TCGCCTGCGC ACGCGTACCT TTGGGGGTTT GGAATGAGCG CGGAAATAAA GAGGCTGTTA 480

ATCTTTTTTT TCGGCGtTCC AACTATTCTT ATGTTGGTAT ATGCGGCACC GCATGcACAC 540

TTCCTAGCGT TCCATTTGGT TATCTTCGGA TCAGTTATGG GTGCGGTATG GGAAATGCAT 600

GCGATGGTGT CgcGcAGGAT GTGCACGTAC CCACTGGTTT TGTTGATCCC TTTCAGTCTT 660

GTGCTTCCGC TTTTAGGATA TGCAGCGCTG TGGCAGCCTG CACGGGGCGC TGAATCTGTC 720

CTTTTTATTG GAGCACTGGG CACGCTGCTC ATGAGTGTTT TTTTCACCGA ATTGGTGTAT 780

TCGTTTTCTG CTTCTTTTGA AAACGCCCTT GAGCGTATGG CCTCGGCACT GTTGCTTGTT 840

TTGTATCCAG GTATCTTTAG CCTTTTTTTT TCGCTCATTA CGCGGTGGCG TCATGCAGAG 900

ATCGCaTTGG TAATTTTTTT yCTCATGGTT TTTACGTGCG ACTCTTGTGC ATGGTTCTGT 960

GGGACGCTCT GGGGAGTCAA CAACAGAGGG ATAATTCCTG CAAtCCTAAA AAGAGTATgC 1020

AGtTTTATgG AGGTTTTgCC GGTTCGGTAG GTGCAGGgTG TTTTGGCTCA CTtGTATTTG 1080

GTTCGCgTGT GACGCTCTCT TTGGGGATGC TCATGGGTGT TGGAGCCTTG GTAGGACTGA 1140

CTGCCATTGT AGGCGATCTA GTCGAGTCGG TGATGAAACG TTCGGCTCAG GTAAAGGATT 1200

CAGGATTTTT TACCCCCGGG CGGGGCGGAA TTATGGATAA CCTGGATTCG tTGCGCCGTC 1260

ACTGGGGACT TTTTACATTG CATGTGAGTG TTTTGGGATC GCTGCAGTAT GAGTGTGCGA 1320

CGTGTGGTAG TGCTGGGCAT TACTGGTTCT ATTGGAGCTG CAGCACTCAA ACTTCTGCGT 1380

CGGTTTCCCG ATCGGTTCTT GCTGGTGGGC GCTTCAGGTC ACCGGCAGAC CGAGTACGCG 1440

CGGGCGTTGG CGCGCGAGTT CTCTTTATCA GATATCACTA TGACTGGCTC ATGTTCTGAG 1500

CAAGAAGGTC GCGCACGCAT AAAGCGTCTG CTTTCTTCCT GTGAAGCAGA GGTGGTGGTA 1560

AACGGTATTG CCGGCGCTGC TGGTCTTTTT GCCTCTCTTG AGGTGCTCAA GACGCGTTGT 1620

ACGCTCGCGT TAGCAAATAA AGAAAGTGTG GTACTTGCAG CTTCTCTTTT GCATGCTGCG 1680

GCACGCGAAA GTGGGGCAAC AATCGTTCCT GTAGATTCAG AGCATGCTGC TATTTTTCAA 1740

CTTATTGCAG cGCACGGCGC GCATGCGGTG GCGCAGGTAG TGCTCAcTGC GTCAGGTGGT 1800

CCATTTAGAA CCTTTTCAAA GGAGTGCTTA GCGCATGTCA CGGTGGAAGA TGCGCTTCAA 1860

CATCCGACGT GGCGTATGGG GAAGAAGATT TCTGTTGATT CTGCAACACT TGCAAATAAG 1920

GCACTGGAAG TTATAGAAGC AGTGCAGTTT TTTCGTATAC CGGTGGATCG GGTCaCGGTG 1980

GTGGTGCaCC CTCAGAGCaT AGTGCATGCg CTGGTGcAAT GTCATTCGGG AGAAACGTAT 2040

GCGCAGCTTT CTGTCCCTGA TATGGCGTCG CCGTTACTGT ATGCGTTGCT GTACCCTGAT 2100

GCGCCTCtGC GTATCAAACT CCGCTTGATT TTACATCGGG ACTGTCTTTG CATTTTGAAC 2160

CTCCGAGGGT AGATGACTTT CCGCTGTTGC GTATGGGTTT TGATGTTGCA CGGGCGCAGC 2220

GTGCGTATCC TATTGCCTTT AATGCAGCAA ATGAGGAGGC GGTGCGTGCG TTCTTGCAAA 2280

GAAACATTGG GTTTTTAGAT ATCGCACACG TGACTGCACA GGCGTTGCAA GAAGATTGGC 2340

GCGCAATTCC CCAAACGTTT GAAgAAGTTA TGGCgTGCGA TAmGCGTGCg CGGATGTGTG 2400

CgCGGACGTG CATTGCACAG AGGTGGAGAG AGAGGTGATT AAGATAATTA TTGGCGTTGT 2460

GGTGCTTGGT ATTGTGGTGT TGTTTCATGA ACTGGGGCAT TTTGTCGCCG CGCTTTGGTG 2520

TCGAGTGGAG GTGCTCAGTT TTTCTGTCGG TATGGGGCCG GTCCTGTTTC GAAAGAAATT 2580

TGGAAAAACG GAATATCGCC TTTCGATGCT TCCTCTTGGG GGGTATTGCG GTATGAAGGG 2640

AGAGCAAGCG TTTCAAACGG CGCTTGATCA AAAACTTTCC CGTATTCCCG TTGAGCCCGG 2700

TTCACTGTAT GCaGTAGGAC CGCTCAAACG CATGGGTATT GCCTTTGCAG GACCGCTGGC 2760

GAATGTGCTT ATGGCGGTAA TGGTATTGGC ATTGGTTAGT GCGCTTGGCT CGCGTGTACA 2820

CACATTTGGA AACCGTATTT CACCGGTGTA TGTATACGAT AGTTCTGATA ACTCGCCTGC 2880

ACGCCGCGTG GGACTTCAGG ACGGGGATAC AATcCTGCGC ATTGGTGACC AGCCGATACG 2940

CTATTTCAGT GATATTCAAA AAATTGTATC ACAGCATGCG CAGCGTGCAT TGCCATTTGT 3000

GATCGAACGG AGGGGGCAGC TTATGCACGT GACCATTACG CCTGATAGAG ATGCGCATAC 3060

TGGCATGGGG AGGGTTGGTA TTTACCATTA CGTACCGCTA GTTGTTGCGG CGGTTGATGC 3120

ACACGGTGCT GCATCGCGGG CAGGTCTTGA ACCTGAAGAT AAAATTCTTG CAGTAGCAGG 3180

ACGCCGTGTG CAACACcAGT ACAGCTCCTT GCGCTGCTCA AGGAATTTCG AAAAAAGTCA 3240

GTCGTATTGA CTGTGCTGCG TTCAGGGAAG AGGCGATATC ATACCATTGC GTTAGTGCGC 3300

ACAGAAAACG GGGCAATAGA TGTTGGTATC GAATGGAAAG CTCACACCGT GGTTATACCG 3360

GGAACTTCTT TTTTTGCAAG TGTCCGTGCG GGCATTGCAG AAACGTTGCG TATGTGTGTA 3420

TTGACGGTGA AGGGTATTGG TATGCTCTTT CGGGGCCTGC AATTTCAGCA GGCTATCTCA 3480

GGCCCATTAA GGATTACGCA TGTGATAGGA GATGTGGCCC AGCATGGTTT TCAGGAGAGT 3540

TTTTTAACGG GACTGTCACA ATTATGCGAG TTTGTGGCAC TCGTGTGCGT CTCTCTCTTT 3600

ATTATGAATC TACTCCCCAT TCCGATCCTG GACGGCGGTT TGATTTTATT CGCATGTGTT 3660

GAATTGTTTA TGCAAAGAAG CATACACCCG CGTGTGTTGT ACTATCTGCA GTTTGTAGGT 3720

TTTGCGTTTG TTGCATTGAT ATTTTTATGT GCGTTTTGGA ACGACGTGAA TTTTTTGTTT 3780

CACTAGGAGT GAGTGATGCA GTTACGGTGT GCGTGTGAGC GGGTGTTCGA TATTGAACAT 3840

GAGACGGTAA TTTCGCTTGA TGAGCACCCG GAATTTGTTG CGCGTATACA GCAGGGGGAT 3900

TTTTTAAGTT ACCAGTGTCC GGCATGTGGT GCGCGTATTC GTGCCGAAAT AAAAACAGAA 3960

TTTGTGTGGC ATGCGAAGAA TGTGCATTTG CTTTTGGTTC CTGaGCGAGA GCGTTTGCGG 4020

TGTTTGGCTT TTTGTGCCGG TATGCATATG AGCGACGGAG ATAGTGCTGA CTTTTGTGAA 4080

CCCTTTGTCT TACGGGAGCA CCAGACACCC GTGATTGGCT ACGCAGAACT TGCTGATCGT 4140

GTTGCAATAC TAGCATGGGA TTTGAACCCT GAAATTGTTG AAGCAGTAAA GTTTTTTGTG 4200

TTGGAAGGGG CACCGCATCT AGGAGACAAG AGAGTTTCGT GTTTTTTTGA ACGTTGTGTC 4260

GGGGACACCG GATCGCGCGT GATGGAGTTG CACGTGTACG GTATCAGAGA ACAACAAACG 4320

GCAATTATGC CGGTTCCCAT GAATGTGTAT GAACGCGTTG AGCGAGAGCr mGGTAAACAA 4380

GCGGAGTTGT TTGAGGCGCT GTATGTTGGG GCGTATCTTT CATACAAGAA TGTTTTTACT 4440

GACGCGTAGC sCCGCACAGC GAGCAGCATC TGGTGTGCGT GGTGTATGGG GTGTTGTACG 4500

CTTGGGCTTT TTGCAGACAG TGTAGAGAAG CGCGCAgcgA AGGATGTGTT TACTGAACCG 4560

GCGCGCTTTT ATCCCTCACA AAAATCAACG CTTGAATCTG CCCGGTCTGA TACATCTGAA 4620

TCTGAGAATG CATCTTCTTC CGTTCCTTCC CACAGTCAGC AGGAGTTGGC GCCAGACTCT 4680

GCCGCGCCTG CGCGTAACTC TGTGTTGTCC CCTGCTCCTC CTGAAAGGAG AGAGAAGCAG 4740

GGGACTGCGG TGCATGGGGC GGAAGTGACG CGGGCGGGAG CTGTCAGCCC GCGTTTTGTA 4800

GGGGGGCTGA CAAAAATACT GGCCGCCTCT GACCATACAT TCTTCGCTGC AGGAAATGAT 4860

GGGTTTCTCA CCCAGTACAC GTATCCGGAT TATAAACCGG ATACGTGGCA GATCACCCCT 4920

GTTTCTATCA AACACTGTGC AGTGCATCCG GACCGCGCGC GTATTGCCGT ATATGAAACA 4980

GATGGACGCA ATTACCACCG AGTCAGTGTG TGGAATTGGC GCACGAAAGA AATACTTTTT 5040

GCAAAGCGTT TTACCGCATC GGTTGTGTCA CTCTCGTGGA TTGTGCAGGG AAGTTTTTTG 5100

AGTGTGGGAA CAGCATCGCG CGAAGgTGTG ACGGTGTTAG ATGGGAGTGG AAATACAGTT 5160

TCTCTATTTT CGGAAGAGCC TGGGGTGGTG TTGTTGACTG CGAGTGGACC GCGCCTTGTG 5220

CTCAGTTATG CAGAATCTGG ACGCCTCACG TACGTAGATT ACAGCAAAAA GACAACCGTC 5280

AAACGTCTTC TTACCGAAAA GAATCTCCTG TCTCCCATGT TAATACATAA CGGTGCACAT 5340

CTTGTCGGTT ATAGAGACCA ACGTGTGTAT GTCATCCAGT CTTCAAGTGG CGCGGTGCTC 5400

ACCGAGTACC CTGCACGGAG tGcATGtTTT GCGCATACAT TCAGCGATAG TCTTCCTGTG 5460

TGGATAGAGC CTGCTGAGTT GAAGTATCAC TGGCGTATAC GGAAAGcTGC GCAGCGTTCT 5520

GCTGATTTTA TGCTTCCTGA CAATGCTCGC ATAACAAGTG CGTGCTCGGT TCGCACGCGG 5580

GTCATCGTAG GAACCGATCG CGGGATCCTC TATGAATTGC AGCAGGGAGA TGACAGGCGC 5640

GTAACTATCC GCGCACTCAA TGGCGAGCGT CAGATATACG CAAGCGATGT ACATGGTGCA 5700

GATGAGGGCG CGTATTTTTT AGCAGACGGA TCCCTATATC ACAGCATGGC GTCCGGGGGA 5760

CCGTATCGTG TTTTGGTGCG CGGAGTAAAA GGAACTCGGT TTCTGCCTTA TCGTGATGGT 5820

TTTATTGTGT GGTCTGCAGG GAAAGAAACA GAGTTTCTTC ATTGTGCGCA AAAGACGAGT 5880

CAACACAGGA TGATATATCG CGCGCGTTCC ACGGTAAGCG GCGTGTCGGT GTATGGGCGT 5940

ATGTTGGTGA TTACTGAACC TTTCTCTGGA GTATCGGTGG TGGATATTGA GCGGGGGATA 6000

CgAGTTTTTT TTCACAAAGC GATTGGTATG CAGGATTCGC TATTGATTAC TGATGACGTA 6060

ATTGTAGCCA CTCAAAGCGG TTTGCAGCCA CTTGTCCTGC TGCATATGCG TACGGGGGAG 6120

ACATATACGC AGCGGTGGGA GGCGATTTGC CTTGGCGTCC GCGCGCATGA TACACAGCAT 6180

GTATATTTTT TTTCGTTGGA TACGAATGCG GGCACGACTG ATTTGATCCA TTTCGTCTGC 6240

AACTGCAGCA ACCCACAGAA AGTGTTGTGC GACGCATCCT CTCTTAtAAG GATGAGGATA 6300

TAGATGCGCA TATGGTGATG CGGCGTTCAC TGTTGGTAAC TAATTTAGGA AAAGGGGCGC 6360

TTGTCGGACA TCGCGTGCAA CAGTCGCAGG TGTATCGTAT GTCCCGTGCG TATGCGTTAC 6420

CAAAAGTTGC TGCAATCACG TCGAACGGAG TTGTCAGCGT GAATTACGAT GGTTCAGTTT 6480

CGTGGTATGA AGGCGACGGT GCGACATTGA AAGCAACCGA ATTTATCCGG ACCGAAGATT 6540

TTTGAACGGG TACACAAGGT GCGGTGTATT TTGTAATTCG GCACGGTGGT ATGAATGCTT 6600

CCTAGTTGGT CTTGACAGGG AGCTCCTTCT CGGGGGAGGA TGGGCGGGGT AGATGTTGGT 6660

TCGCTACAGT TACGATGCAA AGGGAAGGCG GTTGGGGCGT GCGCTGGTGT ACACTGAGTC 6720

GGAGCACGGT ATACCTCGGC AGAGCGTTGA CGCTGGGGCG ATAAGGgTTG TAGAAGCGCT 6780

GGTGGGTGCG GGGTATGAAA CCTATATCGT CGgTGGGGCG gTAAGGGACc TGGTTGCGGG 6840

AAGGACACCA AAAGATTTTG ACATTGTTAC AGGCGCAGTT CCCTCTAGGA TTCGTAGGTT 6900

GTTCAGGAAC TCGCGCATTA TCGGCAGGCG CTTCCGCATT GTTCATGTGT CGTGTGGCTC 6960

GCAGCTGTAC GAGGTTTCCA CCTTTCGCTC TCGTGTGGGG GAAGgTTCGG TGTGTGTTCC 7020

TGGCACGTTG GAGGAAGATG CATGGCGGAG GGACTTTAGT GTCAATGCCT TGTACTATGA 7080

TCCTCTGAGA AATGTGGTGA TCGATTGTGT CGGTGGAATG GTTGATCTGA AGAGGCGTCG 7140

CGTGCGGCCG CTCATACCTC TGCGGTCCAT CTTTGTAGAG GACCCAGTGC GCATGCTCCG 7200

GGCATTGAAG TGCTCGGTGA TGTGCGAGTC TTCCATCCCT TTTTCTGTCC GCCGCAtATT 7260

CGCCGCAtGT TTCCCTTCTt GGGGGGTGCT CTCCCTCCCG GTTGACCGAC GAATTtGTAA 7320

AAATCCTCTT TtCCGGTCGG AGCGcCsCGC TTGTGCGCGC CCTATGTGGG TAmCAGCTCC 7380

TTCTGTACTT GCAGCCGTCT GTGCACTACT TTATG 7415

(2) INFORMATION FOR SEQ ID NO: 6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5271 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double

(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

CTTTGATTGG TAAGGTAAGA GAACCGTCAA GAAATAATCC TTTTAAAGTT TTTAGTTCAA 60

AAATAAAACG CGTTCCACTT ATTTCAATAG TTTTCGTTTC TCCTTCAGTT CTAGAATAGC 120

GAACGTTTCC TGTGGAATAC AGAATTTGGG CAGAGGTGTT GTAGACAGTT TGATCTGAAA 180

GAATCGTTGC GGTAACGcTC CCATCATCAA TGGAAATGGA AACATTCCCG GTAAAGACTA 240

CCAACTGATC TTGCAAATCA AAGGGGGACA GCGGCACGCG gCCGGTACTT TCCAATATAG 300

GTTGTCCAGT TTCAGAAAGG CGCGTAGTTT CCTGTGCCGA ATTAATAATA ATTCTTAATT 360

TGCGTAGACC ACTTTCACCA AAGAGGGGAC AAAAAAGAAA GAATATGCCC CAAATTGGAT 420

ACCATGCTCT CATGTCTTTT TACACAGCAC CTCTTTGATG TACCAACCGG TGCGAGACTC 480

AGTTATCTGG GATACTGCTT CAGGGCTTCC CTGTGCAACG ATAGTTCCAC CGTGCATTCC 540

TCCTTCAGGA CCTAAATCGA TAACACAGTC TGCCTGAACA ATAACATCCA TGTTATGTTC 600

GATCATCACA ACCGTATTTC CCTGATCTAC CAAGCGTTGA ACAACCTCCA TTAATTGGAT 660

GATATCGGCA AAATGCAATC CGGTAGTAGG TTCGTCAAAG ATATAGAGAG TTTTTCCTGT 720

CGCACGCTTT GAAAGCTCAA GTGCAAGTTT AACGCGCTGG GCTTCTCCCC CTGACAACGT 780

CAGAGcAGAC TGTCCTAAGC GCACATACCC AAGCCCCACC GAGcAGAGAG CTTCTAGCTT 840

TCGTACTATA GGrGGaACAG CAGAAAAAAA AgAACGsGCT TCTTCGATCG TCATGTCCAG 900

CACATGGGAA ATGTTCTTGc CCTTATAAAA CACAGCTAAT GTCTCCCGGT TAAACCGGGT 960

GCCGTGACAC ACATCACAGG TAATGTACAC ATCAGGTAAA AAATTCATTT CAATAGTGAT 1020

AACGCCATCA CCTTTACAAT GCTCACACCG TCCTCCAGGA ACATTGAAAG AAAAACGTCC 1080

TGGTTTATAT CCCCGCATTT TTGCTTCAGG AACcTGGGAG AACaGCATTC TAaTATCTGT 1140

AAACACACCC ACATAaGTTG CAGGATTTGA AcGAGGAGTT CTCCCGATAG GACTTTGGTC 1200

TACATAAATT ACTTTATCTA AATGCTCCGT CCCCTCAATC GAGGAAAATT TTCCTTcAGG 1260

AAGTCTGCCG TTCATCACAC GGTTGTATAA CGCAGGATAT AGCACATCAA TTAAAAGCGT 1320

TGATTTACCC GAGCCGGATA CTCCGGTAAT GCAGGTAAAA GTACCGAGTC GAATACGTAC 1380

AGAAATGTGT TGCAAGTTAT GTTCATGGAC GTCATGCACC GTAAGAACAT TTCCATTTCC 1440

CGTTCTCCGT ACTGCAGGAA TGGGTAATGT AATTGCACCG GCAAGATACT GACCAGTAAG 1500

ACTTGCTTGC ACCTGCATAA CTTCAGGTGG ACykCtGCGG CGACAACATA TCCTCCGTGA 1560

ACACCCGCAC CGGGGCCGAG ATCTACAATA TAATCTGCTA CGCGGAGcgT TTgCTCATCG 1620

TGCTCTACCA CAAGCACTGT GTTTCCCAAA TCACGCAAAT GAAGAAGCGT TTGGATCAGT 1680

CGTTCATTAT CCCGCTGATG CAAACCAATA GACGGTTCGT CCAGTATGTA CAAAACCCCT 1740

GTAAGGCGCG AACCTATCTG GGTTGCCAGT CTAATTCGTT GTGCTTCTCC GCCGGATAAC 1800

GTGGCAGCAG CCCGTTCCAA GGTGAGATAT CCAAGACCCA CGTTCTGAAG AAACTCTAGG 1860

CGATCGGTAA TTTCTTTCAG GATCTGTTGC GCAATTGTCG CTTCTACTTC TGTCAGATGG 1920

AGAGTTTTAA AAAACTCACA CGAATCATCT ACAGACAGCg CACTGAGTGC GTGGATGTTT 1980

TTTTTTTCTA TAGTCACCGC AAGCGACTCT GGCTTTAAGC GCATCCCTCG ACACGCTTCA 2040

CATGTACGCA CCGATAAATA CCGTTCATAT ACCTCGCGCT GTGAGTGAGT ACATGACTCT 2100

GCGTATCTCC TGTGCAGCTC GCTAAAAATT CCCGGCCACG GCTTAATGTA GCGTGCGGTA 2160

CGAGAGCCAT CTTTTCGTTC ATGGGAAAAC TCAAGAGCCT CGCTGCCACT TCCATGCAGG 2220

ATAATATCCA GTGCGTGTTT TGACAGATTG CGTACCGGAT CATCGAGAGA AAAATGGTAC 2280

TTTTCTGCGA GTGCAGCAAA CCGCACACGG TTCCACTCAT GCTCAGGTTT AAATGGCAAA 2340

AAAGCACCCT CGTTAAAAGA ACGGTTTTGA TCAGGGACAA TGCGATCTAA ATCAAATGTC 2400

TGCATAATCC CCAGTCCTGC ACAGCTCGGA CAGGCACCAA AAGGTGCGTT AAAAGAGAAC 2460

AAGCGAGGCT GCAATTCGGG TACGGAGACA TTACAGTGCG CGCACGCGTT TTTTTGCGAA 2520

AAAAATAACT CAGACGGCAG GAGAGCAGAT GTTTCTATCT TTCCGGAAAC GGTCCCAGAA 2580

TTCTCTCCCT GCACTAAGAC GGTCAACAGC CCATCTGCAT AGCCTAGCGT CGTCTCTACT 2640

GATTCTGTTA ATCGTTTACG TACTGTATCT GACAATTGAA TTCTATCGAC AACTATATCG 2700

ATAGAATGCT TTTTTTGCTT ATCCAACGAA ATGCGCTCGT GTAAGTGGAG CAAAGCCCCG 2760

TCAATACGAG CTCGTACAAA ACCATCTTTG CGTGCAGCTT CCAAGACCTT GTGGTGTGTA 2820

CCTTTTTTTC CTCGCACCAC CGGGGCAAGC AACTGAATTC TGCTTCCCGA CGGCACGGTC 2880

ATGAGGGTAT CAACAATTtG GTCAACGGTT TGTTCCTtGA TCTCCCGCGC ACAGTGCGGA 2940

CAATGCGCGC GTCCTATGCG GGCAAACAGC AGACGATAGT AGTCATAAAT TTCTGTGACA 3000

GTACCAACCG TTGAGCGAGG GTTACGCTGc GTAGTTTTTT GCTCGATGGC AATCGCAGGA 3060

GAAAGACCCT CGATAGAGTC AACATCCGGC TTATCTAACC GACCTAAAAA CTGGCGAGCG 3120

TATGCAGAAA GGGACTCCAC ATACCGACGC TGTCCTTCTG CAAAAATAGT ATCAAACGCA 3180

AGCGAACTCT TGCCTGAACC AGAAAGACCG GAGATCACCA CAAGCGCATC TCGCGGCAAC 3240

ATAACATCAA TATTCTTCAG ATTATGCTCA CGCGCACCCT TTATACACAG ATTACGAGCA 3300

GCAGAGCCCA CGCGAACGTC CTGCGACACG TTCTCCCTTT CTACGCGATC CCCATCCATA 3360

AGAGGGCAGA CTATGCGCAA TTTTTATGCT TTATGCAATT CCACTCCCTT CGCGGCAGAC 3420

CGCATTACAC CGTTCCTCCA AGCAACGCGA CAATTTTTTC ACTCTGCGCA GACTGTACGA 3480

TGAACATCAG CACACTTCCC TCAAGCAGAA CCGTATCTCC TGAAGGAATG AAAGAGCCAC 3540

GCACCGTTGA GATGAGCAGC ACCAAGAAGC TTCCGTGCAC AGCGATATCC TTCAGGCGCT 3600

TGCCTACCAG GGGGGACTGC GCGGAAATAG CGAACTCAAC GATTTCTAGC GATCCGTCGC 3660

CGATAGTATG TATGCCAGTG ACGTGGGAAC CGGCCAAATG GCTCATAATA GCGTCAACCA 3720

cTACGTCTTG ATAAGAAACA GCAACATCGA TGCCAATTTT CCCCGCAATA TCCTCCATAA 3780

GGGAGCTGTG TACCAATGCA ACAGCCCGAG GCACTCCGAG CGTCTTCATG TATGCGGCTG 3840

TAATCATATT CAGCTCATAG TTATTAGTGG TGGTAATCAC CAGATCAAAC GTGTCCGGCG 3900

TAATCTCTGC GAAAAAAGCC TCATCTGTGA CATCACCATG ATAGGCAGTA ACGTGCGGAA 3960

ATTGAGCACA CACTGCCTGG GTTGcCtTTC ACTCTTATCA ACCAATACAA GACTCGCACG 4020

CTCCCTTGGA GAAAGACTGA AGGCACTACT GAAAAAGTGC GGCTTGCATT TTTCTGCTAC 4080

ATCCTGTGCC ACGAGCGTAC CTACCGCGCT CATGCCAATG AGTGCAATTT TTTTTACCGG 4140

ATGTATTTTA AAACCCGCCA GCTCATAAAA ACGTCCCATA TGTTCAGGCG CACAGAGTAC 4200

TGACAGGCGC ATACCAGAAG CGAGCATGGT CTCCCCTGAG GGAATTACAC TCCTCCCCCG 4260

AACTTCAAAA GCAACGGCAA CAAAAGAAAT TTTTACCAGA CGACGCATAT CAGAGAGCGT 4320

GATACCATCG AGGCCGCTGC CCTTTGCAAT AGGAAAACGG GCAATTTCAT ACGGTGCATT 4380

TTTCAATGGG ATGACATCGC TGATGGCACC CTGCTCGACG GTGCTCACTA CCGCACGCAT 4440

CGCTTCCTTA TCCgCAGATA TGAGAAAGTC AATACCAAAA ATACAGCGCG ACTCACGACA 4500

CACCGCGTGA GcGTAGTGGT CATCGTGCGT TTGGGCTATT TTAATCACTC CGGCATTCAA 4560

GTCGGCGGCT ATACCACACA GTACTATGTT AAGTTCGTCA ACCTCGGTGA CCGCAACAAA 4620

CGCCTGTGCC TTTGCGATAC CTGCyTCACC CAGGGTAGCg CGGTGAtCTT TTTGATGACG 4680

CaCGAGCGCT TGTGCCCCCG CGGGAATTTT ACGGGGGACA GGCTCATGCG CGGCAGCAAC 4740

AAGCGTAACC TGATGTCCCC TCGCGCTCAA ACGACGCGTA AGTTCACGCC CATTCGTACC 4800

GCATCCAACA ACAATAACCC TCATGTCCGA AGGCCATAGT AGCACGAAAT TTTTTTGCAT 4860

GGCCAGCGCG cAGAACaCGg CGcACAACGC CTGCCACTCA TATCTTTTTC AAAAGTACCA 4920

CTACCTGTGC GGTAACCGCC GCACCAGATC CAACAGGTCC GAGGCGTTCG GCAGTCTTTG 4980

CCTTAACAAA AACACGTGTT ACGTGCGTGT CCAGGGCCTG CGycAaGcGA TGCGCGCATC 5040

GCTTCCCGAA ATGGGTGTAA TGCAGGCTGC TCAAGACAGA CAACAGCATC GAGATTCACC 5100

AGCCGCCAGn CACTGCGCGC ACCAGTTGCC AGGTATGGCG GAGCAACGCG CAAGAATGTG 5160

CGTCTTTCCA TCGTCCGTCA CAAGAGGGGA AAAACGTGCC AATATCCCCC AGGCCCTTTC 5220

GCCAGCTGGC GTAATAGCGA AGAGGCCCGC ACCGATCGCC CTTCCCAACA G 5271

(2) INFORMATION FOR SEQ ID NO: 7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 646 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:

AAGTCCTTCT CGTAGCGCTT ACTGTCGGTG AAGCGGGAGT GGAGGACTTG GCCCGGCAGA 60

TGCAGCAGCC CTCTATGCTC CGGTACATCT TTGGCTTGTG GAAACCAAAC ATCTTCGCTT 120

CTGGCCAGGT CTGGAAAGAC AAGGCAACAG ACACAGCTCT GCCGAGGCCC TCACAAGAAA 180

CCATCCGGGG AGCCTGTGCT GCTCCACCCC GGGCCCACCC AGCGCCGTCA GACAGCAGGA 240

CTCTGTGGAA GGTGGCCTGG ACCCGCCTCC GCTCCTGGCG CTGGCACGGC AAGTGTATGA 300

CACACAGAAG AGCTCAGGTG TTCAGGGAGG CCCCCCGCTC TCAGCACTCC CCCACCCCTG 360

CCCAGCAAAC ATCCTTTCTG AAAATGAGGA AGGGGAGGCT GGTTGGTTTG TTGGCAGGGA 420

GCCAAGCACT TGAGCCATCA TCTGCTGCCT CCCAGGGTCC ACGTGAGAAG GAAGCTGGAA 480

TCGGGAGTGG ATCAAGGATT GGAACCCAGG CACTTGCGTA CAGGATATGC TACAAGCTCT 540

CCTGATAATC CTGTAAAATG ATGAAATCAT TTAGGATGTA TCCTGAAATC TGAGACaAGG 600

CATACCTTTT CTTCTTGCAT CTTTGAAAGT GaACCCCCCC CCACGC 646

(2) INFORMATION FOR SEQ ID NO: 8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 28295 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 8:

GTTCCCTAAA GAAGGGAATG TTTTCTCCtG TGTGTGCAgT CAATGTGCCG GCGTATATyC 60

CGTTAGAATG GGTGCGCTCC AGTGTCAAGG CGATATCCTG ATAGCGCATT CTGAGCGCTC 120

GGATATTTTT TTTAAACACA TTAGTCATAT GCTGAGCTTG TTCCTAAGTT GAACGATTTC 180

ATGGCGCACC TTGTGAATGA GAGTTTGTGT TCCCTGAGAA GGGTTTTTTC TCAGGTGAAG 240

TAGCTCAGCG TATGCAAGAA ACTCAAAAGC CTCTTTTTCT ATGCCGGAGC TGGCTATACT 300

GGATGCGATA CGTGCATGGT GTGsACTGCG TGTGATGCGT TCAAAGTAGT GGCGGAgCAC 360

TGCCTTTTTG TGGATGCGCG AAATCGGCGG CCCTTCTGAA AAGGTGAATG TATCCTTTTT 420

GGCGAAGGCG TGTTGGGAGA AATCTTCATA CGTGATGGTA CGTACATTGG AAGGTGGTGC 480

GTGTTCTGTA TTTAGCCGTG TAAGTTTCTG GATTTtCTCT TGTGGGAGAC TGTAAAACCA 540

GTGTGCGTAT GTTTCTAGTG AGCGATTGTC AAAATTTCCT GCAATGCAGt CGCAAGTGGC 600

TCAGTTCTCT TGGGATTAAA ACGTTTGAAG GAAGCATGTG GACGTGCGTG TTGGAATCCC 660

TTTCCCGGGT GCAAATCGAG TCCGGCCAAA TAAATAGAAT GTACCCCACA CTGTAAAAGA 720

TATTCGAGTG CAGTCCCCAT GACACTCCCA TGTCTTTGTG CAGAAACCGA AGGaTGTGCA 780

GGTGTTGCAG AAAGAAGTGT TCTGTATGCG AATGGTAGTT GAGCAAAGAA ATTGGAGAGT 840

ATTTAAACAC ACGTGGAGGA ATGCACGCTT CTAAGGGAAA CAACACCGGA AGGGTAGGCG 900

CTGGCGGAAA ATGCTCTGCG GCCCAAAAAC TGCCGTCCGT GCTCATGCAG ATATCAGGAG 960

AAATATTGCG GTAGAGAAGT GCCTGCAGTG CTGAGGAGAC CGCAACTATG GGAAGTCCGA 1020

CTCGGGTGAT TTGCTCGAGT CCGAAACCTG cAGCCACACA CAACACTTCC GGTGCGGTCA 1080

GGCACACCgT ACCGGGcGCT CAAGAAAAAA TACATTTCTC AGTGTATTCA GTAACCAACG 1140

TTTTCCGAAG TATATGCGCG TGGCGATTTC ACTCTGAATT ACTTTGATGG TATAgGTGAT 1200

TTCTTGCCAT GTGGACTGAG CTTGTGCGCG CCATATTCGG TTTGCTGGCT CCCAGGGAAC 1260

AAACGCGGTT TGCCCGAGCA ATTCGTCAGG GATATGATTA ATTAAGAAAG AATGCAAACT 1320

GCcGCAGTCT GGTCTCCAGA CTGCGTCCCA CTGGATATCG GATGAAGTGA ACGCATCGTT 1380

TGTGTATCGA ACTGCGATAA GCTTTGCATG TGGAAAGCGT GCCCGCAAAA ACTCCGCGGT 1440

GTACGACTCA CCCGGCTCTG TTATTACCAC AATGCGGCGA CCTTCTAGGA TCGCAGCGAC 1500

AAAACGCTCC GCCTCCCTTC TTGGGTTATA TTTTGAGTGG AGGTTCAGCG ACACGTCTCC 1560

CCGGTTTCGT ACTGCAAGAA ACGGAGGACG TCCGTCTGCG CGCGAGTAGT TCCACTGCTT 1620

GTGATGAGAA GTGAATTCTC CGCGTGGTTA ACAAAGTACG CGCTGACGCC ACGGGTAAGT 1680

TGGTTTAGGT ACAGCTCAGG ATCGACGGGA AAGGAAGAAA AGACCACAAC GCGCActTCC 1740

CGGGAAGCGA CGCATGAGGC AAGGTTGTCG TTCCCTACGA GCGTTAATAG CTGGTTGCTC 1800

AGCGCACAGC AGCGGCGTAC CATACTCAGC TCTTCCTGAA GAACGGCGGG GGGCAACGCA 1860

CTcTGGTTGA TGCTGATAGT GAACAGGTTT GCCATCCGCT GATTGTTCAG GGCGAACTCG 1920

AGCTGGGAGC AGCTCGGTGC CAACTCCGGC TGTcGCGGAT ACACGGGCCT ACTCGCCTCG 1980

ATAAACGAGA AGCGCGCCTG TACGGCaAGG ATAAAGGAAC GCACGCGCTC CTCGCCTGAT 2040

GCaTCACCGT GAAGGTCTGC aGCGAGGGCA CCTTCCTCGA CGTCGGTTCG GGTATCAAAG 2100

AGCACTAGGT ATACGCGCAC GCTTCCAGCG GAAAACGGAT ACAAGTGGAG CGCACGCAGC 2160

GTGCTGAATA GATGGGAAGA GAAGTAACCC TGCAATTCGG TAAGTTCCTG CCTGAACAGG 2220

GTCGTCCACT GCTCCGCAGg CAGGCCGGGC GAAAGGAGGG AGGTGGGAAC ACGCATACGG 2280

TAAAACGTGG TGGTGTCAAT CCCGAAGGGa AACGCAGgTC AAACATGCGC TCTTCAGTGA 2340

GAAAGAGAAA ACCCGCGCGG GGAATTCCGA GAGATTCAAC CAGCTGTACG AACTGCAGCT 2400

CCAGGCGGTA GTACTCGCAG GAGACACCCG AGCTACGCCG GATGCGGGAA GCGCGCGCGA 2460

TAAGACCCAT TAGGAAATCC CTAGCTcGTC AAAGAGGTGT TTGTACGTGT GGAAATACTC 2520

CGATCGCGCA AACTCCTCGA TCTTCTCCTC GGGCAGACTT TCGAGCAGCC GGTCCATATA 2580

CCCGAGCACA GAGCGCACCT CGTCGGCAAG GCTCTTTGAG AGGGGAGGTG CAAGGTTGGC 2640

AGAATCCCCC TGCGCAGGAC GGGCCGCTTC CGATTGGGGC GGGGACTCGA TTGCCACCGG 2700

CGTGGCAACC TCCGAAGACT CAGAGGCGGC GGTTTCAATC GCGTGGGGAG CAGAAAACTC 2760

TAGCGAGGGA ACGGAGACGT CCAACTCCTG CTCTTCTGGC AGTGGGAACG ATGTATCTTC 2820

GCGCTGGGTG TCTTCGTCGT CGAAGACATT TTCAGTATTG AGGGGCTGCT CGGCACCGGC 2880

TGCATCGAGG CCGGTGGAGG AAACGTGCGC CCGTTGAGCC TCAGGCGATG CGTCATCTTT 2940

GGCAACAGGA GACGGATCGG CCGCAGCGTC ACGATCCGCG TGAGCCCCCT GCTCTTTCAC 3000

ATCGAAAAnT TCGTTTCCAA GAGTAGCACG CCCCTGTTGT ACCCACTGCT CCtGcgCGAC 3060

GCgCGCTGcA GCGTGGTCGT CCTCCTCTAG GTAGCTGAAA TCGCCTGCAG CGGACCCGAT 3120

GTGAGAGGAC TCAAAACCAC TGAGTTCACC ACCAAATGCG TCGGCCGCAG AGTCTTTCCC 3180

ATCTTCCTCG GTAAACTCAG AGGTGATGAG GATATTGTTC AGCTCATCGT TGGTAAGCGC 3240

AATCGTCTCG TCTGGATCAT CGTCACAGAA AAACCCGGAG TAGGCAGCCT CTGTTCCTTC 3300

CGGAGCCGGA GCGGTGGCGG GAGCCTGAGC AGGCTCGGCG CCAGCTTGCC GGGAGAAGGT 3360

TCCCTTCAGC TGGTCTAGAT CAGCGCGGAT ACTGCTGATT TcCTGTGCAA TTTTCAGAAG 3420

CAAATCGGTG GAAGCATCGC CTGaGGTACT GGCTGCGCCC CTGCGCGGCC ACCTGGGCAT 3480

CACCCATGAG GGACTGCGCa ATGCGTCCAC GTCACTGAAT GGGGCGCCGC GTACCGGTCC 3540

TTCCGAAGAA GGCTCCGAAG CaGaGTAArT GATAGGGTCG GAATCGACCG GGTATTCCGG 3600

TGCGTCGTGG CTCAAACTCG AGAGCAAGTC GTCGAACTCT GTGGTGTCGC ACGACCCGCT 3660

GGGGCAGGGC ATGCTGCACC CGCGCGGCTC TGGCGGCTCT TCGGTCACAA CGGAGCTCAT 3720

CGCTGCCGAT AGATCAACTC TGTCGTCCTG AGACTGCGCG GTGCGTGTGA CCAAATCGTC 3780

GAACGACGGC ACCTCCGAGT AAGCGGCATC TCCGGtGGCG CATCTGCCGC CTGGTCGGAG 3840

GAGGACTCGC CACCCGGAGC GTGGCTGCAC CCGCGAGCGA GCACCTCGTC CTCCACGGAG 3900

AGGTCAACAA AGCCACGAGA ATCCTCTGCA GAGGACTGCA CAGGAAGCGC ACGGTACACA 3960

TGGTGCCGCA CCGCATCCAC GCGCGCACCC GGGGGAGCGT GTTGGGCAGG CGCCTCTGGA 4020

TCCTCTGGGC CACCTGAGCT GACGGGTGGC TCCTGCGCTT CCGGCGGTAC CTTTACCCAC 4080

ACGCCACACG CATCCCAGGC ACGGTCTTCC CCCTGCTCCT GAGGGCGTGC AGTTGCGTGT 4140

GGACTGTCCA TTTCTGCTGA CTCTGATCCC ATAGCATCTT CCTCTGTGCC GTGAACTCCC 4200

TCACACATCC TGTGTCGGCA GCCGTGCCGC AAAGCATGAG CATAGCGTGT GCGGTCCGTT 4260

TTTTCAACAA TTCGTAAAAG CACCCCGTTC TTACGCCGTA AATGACGGCA ACGCGCGGGT 4320

GCGGAAGGCA CCTGCAGTGC CTGTGTCATT ACCTCCTGGC GGGGGGGGGT ACGGACAAAA 4380

ACAGGGGTGA AGATGTGGAA GACAGTGCAG ACACGCAGTC AGAGGAAGCG GTGAGGACTC 4440

TATGCGCATT TACTTGAGGG TAGTACTTCC CCTGTCTCTT GCGCTGAACA GCTACGGTGT 4500

ACTCGCCTTT TTCTGGGGAG AGCGGGGGGT GTGTGCCATG CGGCTACTGG AACGTGAGAA 4560

AAAGGAGCTC GTCCATCACA TCCAGACGCT CGCAGAGCGT GGGCGCGACT TGGCTGCGGT 4620

GGTGGACGCT CTATCCTTTG ACGAAGAGAC TATCGGTGCG TATGCGCGTc AGCTAGGATA 4680

TGTCCGCGCG GGGGATGTGT TAGTGAGGCC GGTAAACTTT ACCnTTGCGC ACATGCATAC 4740

CCTgATTCTG GGGATGCACG TCCGCTtGTT GCACCTGCAT GTTTTAGCGA CACGCGTTGA 4800

CAAGGTGTAC GCGCTGTkCs TcGGCTTcTT TgTCGTCCTG TTACAGCTGC TGTGGGGTAG 4860

CGCGCGTGCG TATTTTAAAA CATGAGGCGC AGTCTGCaCG CTnTGCgCgT GCGCTCAAGG 4920

CCGGTGCGCT CGTAGCGTTG CCGACAGATA CGGTGTACGG TTTCTCTGGC CTTGTGCCAC 4980

ACGCTGTTCC GGATCTCATA TGTCTGAAGG CGCGTGGGTG CACAGAGACG GAAGGGAACC 5040

GGAGAGAGGG CTATCCGTTC ATTGCACTGC TTGCAGATCC ACAGGACGTG GTTGTCTATA 5100

CCGGGACGCG GCTTCTGCGG AGTTTCGTGC GCTGTGGCCT GGCCCGTATA CGTTTGTntG 5160

CGCATGCAAG ACGGCGCGAC GCAGGcgTTC CGCTGTCCTG CTGACCTGTG cTGCGCTCAG 5220

TGATACGGGC AGTTGGGGGA GCGATCTTTT CCACGAGTGC AAATCGGCAC GGCGAGCCGC 5280

CGCTGCAAGA TGCACAGGAC ATCGACCACA TCTTTGGAAA GCATCTTGCG CTGACCGTAG 5340

ACGCAGGACC ACTGACCGGC TCCCCAAGCG CGGTGATAGA CCTCACGCAC CCCGTGcCGC 5400

GTGTGctCCG CGCTGGTGCG GCGCCGTTGC CTCTTGCAGG ACTGGAAAGG CGTGACTCTC 5460

CTTCCCTCCC TCATGTGGGG GAAGTATGTA AAGAATGAGG TCAGTTGCGG ACCCACTCCT 5520

CAGGCACCTC TTCATACTCA ACAGAAGCGG TGTGCCCTGC GCCCCGCACG ACAGGGGCGG 5580

TTCTGACGCC TGAAAGCCGG TTGTTTCCTG AGATCTTTGG GAACCTCCAC TGGAGCGGAC 5640

GCGCCTCcTG CGCGAGCTTC TGAGCGATCT CAGGGGGAAA ATGAGAAAAC TGTCGCGCAT 5700

TGACTGCCCC CTTCTGTGAC ACCACGAGCA CTCCGTCTGA AATTTTCAGC TTGAGGGGAA 5760

CGGAAAGTCC GGAGCGCAAC ACcGCGATGC CGCGGCCTTC GCTCACGATG ACGATTCTTT 5820

CCCCTTCCTC TGTGCTGTGC CAGGAACCGG CGAGGGCATC TAGGGACGAA ACGGATTCTT 5880

CACCTGCGCG CGTGTGCCGC ATGCTAGACG TCTCTGTCTG ATTTCCTGTC AGCGGGACAG 5940

AACGGTCAAA CACATCTCGG ACCAGGTGGC GCGAGTCAAG CAGAATGCGC GCTGCCGTCT 6000

CATATGTTTT GGAGAGCAGC CGCGTGGCGT TGTGATCTTT AmCCTTGAGT GCAACAGCCA 6060

GTCTAATCCC CTCAGGGGTG ArGTCCATCG CGCCGCAAAA TATGTAGTCA AGATTCCCTT 6120

TTTCGGGAAA ACGGTGCGGC ACCGCTTGCT CTCTGCAATC TACAAmGTGA TACCCACGCA 6180

ACTCCCGAAT GAAAgAAAAG AGCGCGTCGT TGATGGTAGT TTCTGTGTGC GCAGGCACAC 6240

CAgACACTTc TAGCCTGTAG ACGCCAACAC GAGGAGCTGC GTGAACAGCA TGCGCGAACA 6300

GCACGAACAG GATAAGCAGC CGAGAAGAAC GGACACCTTT TTTCATGAGA CTAGTGGTGT 6360

CGCTCACAGA GGCTGCGGGA CAGCTCCCGT GCGTTGTCGC GAGCTTTGAT CTGCGCGCGC 6420

TTGTCAAAAA GCTTCTTGCC CTTGCAGATT CCCAGCGCTA CCTTCACCCG CCCTGCTTTT 6480

AGGTAAAACT CCAGGGGGAC CAGAGTATAG CCTTTCTCTT CAACCTTGCG CTTCAAGCGC 6540

GCAATCTGGT CCCGATGTGC CAGTAACTTC CGCATCCGAT CCGGATTGGG GGCAAAGGAG 6600

CAAGCATGCA CGTACTCCGC AATATGCACA TTCTTTAGCC ACAGCTCGCC TCCGCGCATC 6660

TCTGcAAATG CGTCAGGAAA AGAAAGATGC CCCGCGCGCA CAGACTTCAC CTCCGTGCCT 6720

TCAAGCGCGA TGCCACACTC TAGACGGTCT TCCACATGGT AATTGAAAAA AGCcTTGCGG 6780

TTCTTTGCAA TGAGATGGGT TCCTGTGCCC CTCATGGCGC CGGATGCTAC CGGATAGGCA 6840

CTTCCCTTGT CAATTCGATT ATCGCCGTGT TAGGCTGCCG TGTCTGGGAG GGACGCCGTT 6900

TTATGTTTGC GCGGTGGAGA AGGTACTCAT ATTTGGCGCG GCGCGAAGCA CGGCGGAATG 6960

CGACCGCAGT TTGTAGTgCT GGGGTGGGCT TCTTTCTGTT CTATCTTTTT ATCACTACGC 7020

ATGTGGTTGC AGCGTATCGC ATTCAGgCGG ACTCGATGCA GCCGACCCTG AGCGCAGGGG 7080

ATTGCGTTCT TGCCTCGTCC CTGTTTCGCT TTGCCCGCAT CAAGCGGGGG GATTTGGTGC 7140

TTGCAACTCC CCTTGAGAAA GAGGATATAG GCCTGTTTAA AAGGGCGATG AATGCTGTGT 7200

TnAGgnTTCG CAAGCCTTCA ATTGTACCGG CCGTTTGGCG CGGCAGATCG CATGTTTTCG 7260

CGGCCGCAAA TGCGCAGGGT GGTGGGCCTT CCAGGGGACA CTGTCTATAT GCGCGATTTT 7320

GTGCTGTACG TTAAGCCCCA CGGTCAGCAA CACTTCCTCA CGGAATTTGA AGTGAGTGCA 7380

GTTAGCTACG ACGTGCGTAA GGGGGTGCTT CCTGAGCATT GGTCTGAACG GCTTCCCTTT 7440

TCTGGTTTCA TGGAAGAGAT GCAGTTGGAC GAGCACTCCT ATTCGTGCTG TGCGATAATC 7500

GAATTGTCTC CAGTGATTCT CGTCTGTGGG GTGCCATCGA CGGTAGTACG CAGATAAAAG 7560

CAAAGGCATT CATGCGTTAT TTCCCTTTCG GAGCATTTGG TGTCTTGTAG TGTGTAGGCG 7620

CCGCATTTGT GGTGCGTGtG CGCATCGTGC TGTTCCTTTT ATCATGTCTT CTGAGGTCGG 7680

TGCGTCTTTG TACGTGCACA TCCCCTTCTG TGCGCAACGC TGTGCTTACT GCGATTTTTA 7740

CTCCCTGGTG CGTTCAACCT ATTTTAGGCC TCATCAGCCT TGTCCGCATT TTATCGATCG 7800

GCTGCTACAG GATGTGGCAT TGCAGCGGGA GTGCTTTGGG GTCCAGGGkT GGCAGACAGT 7860

GTATATGGGT GGAGgTACCC CTTCGCTATT GGCACCGCAG GACATTCGTC ATTTTTGCGT 7920

AGCGTTACGC GCCGCGCAGC GGTATCCGAT TCAGGAGTTC ACTCTTGAGG TGAATCCTGA 7980

GGATGTGACC GAAGAGTTTT TGTGTGCGTG TGCAGAAGGC GGAGTAAACC GTTTATCCCT 8040

TGGGGTACAA AGTCTGCGTG ATGAGGTGTT GCGTGCGGAG CGTCGTGCAG CCTCTGCTGA 8100

ATGTGCTCGT ACCCGcTCCG CGTGATGACG GCAAATGCGC GCTTTTTCTC TGGCGGGGTG 8160

CGTATTTCAG CAGATCTCAT CGCTGGATTG CGCGGGCAAA CGGCGCGAAT GGTGCGTGAG 8220

GATaTAGATG AGCTTTTGTC TTTTGGGCTG AGACACGTGT CGCTATATGG GTTGTGTGTA 8280

CCGCATCCGA CTGAAACGCA AGAGGAGCGA ATTGCAGCGC TTTGGGCACA CGGCAGCGCG 8340

TATCTGGTGC GTGCaGGATT TAACCGGTAT GAGCTTTCGA ATTTTGCACG TACTGCgGCG 8400

GACGAGAGCG CGCACAACAG AGCATATTGG CGGATGGCAC CGCACGCAGG GGTGGGGCCT 8460

GGCGCAGTTG GCACGCGTTT TGTCAACCTT TCTTTATCAA AGGAGGGGGC GTGGGCGATC 8520

CGCAGCACGG TGCGGAAACA TCTTGGCCAA TACTTAGCAG AAGTGTGTCG GGAAAATGTG 8580

TATGAGCACG AATTCCTTAC AGAACATATG TGTGTGCAAG AAGCATTGTT AATGGGATTA 8640

CGTCTTGAAC AGGGACTGGA TGTGGTTACA TTTCGTGCGC GGTTCGGGAA GGGAATTGAA 8700

GCGTACATTG GCAAAACAAT CGCGCGGTGG CAGTGTCATG GCCGAATGCA GCGGACGGCG 8760

ACGTCATTGC GTTTGAGTGC GCAGgCACGG GTATTTCTGG ACAGTTTTTT GCGAGAGGCG 8820